Handbook series ON electromagnetic interference and compatibility». Перевод с англ. УДК. 621.391.828. Москва. 31. Ослабление стоячей волны в лаборатории радиопомех. W. E. Cory, W. C. Dolle, F. C. Milstead, Standing wave reduction in an RFI laboratory, IEEE transaction on EMC, V. EMC–7, № 1, march 1965, p.p. 64–72. 32. Яковлев А. Д. «Химия и технология лакокрасочных покрытий», Ленинград, «Химия», 1981 г. 33. Рейдман А. И. Защитные лакокрасочные покрытия, Л., "Химия" 1982. 34. Экранирующая оболочка из электропроводных элементов. Заявка № OS 3620172 ФРГ, МКИ4 H05K 9/00, H01R 4/62, H01R 11/09; E04B 1/92. Заявл. 14. 06. 86; Опубл. 17. 12. 87. 35. Способ соединения электромагнитных экранов. Заявка № 62–43560 Япония МКИ4 H05K 9/00; B23K 1/14. Заявитель «ТДК К.К.» Заявл. 02. 02. 81, Опубл. 14. 09. 87. № 7 – 1089. 36. Проводящая прокладка с огнестойким и износостойким покрытием. Conductive gasket with flame and abrasion resistant conductive coating: Патент 5045635 США, МКИ5 H05K 9/00, H01S 4/00. «Schlegel Corp.» – № 397210; Заявл. 16. 06. 89; Опубл. 03. 09. 91; НКИ 174/35 GC. 37. Уплотняющее устройство с электромагнитным экранированием. Заявка 3742762 ФРГ, МКИ4 H05K 9/00, F16J 15/08, Philips Patentverwaltung GmbH. – № 37427628; Заявл. 17. 12. 87, Опубл. 29. 06. 89. 38. Способ и устройство для экранированного соединения между соседними блоками типа экранированной комнаты или камеры. Заявка 131500 Япония, МКИ 4 H05K 9/00, E04H 9/14 «Сумитомо суриэму к.к.» № 62 – 186803; Заявл. 28. 07. 87, Опубл. 01. 02. 89. Кокай токхё кохо. Сер. 7(2). – 1989. – 30. – с. 499–502. 39. Скрепление для механического и электрического соединения тонких экранирующих стальных пластин. Заявка 281499 Япония, МКИ5 H05K 9/90, «К.К. Хитати сэйсакусё», «Хитати биэо эндзиниаринг К.К.» № 63–230020; Заявл. 16. 09. 88; Опубл. 22. 03. 90. 40. Дверное уплотнение, служащее электромагнитным экраном. Заявка № 0269206, ЕПВ (ЕР), МКИ4 H05K 9/00, «The Marconi Company Limited». Заявл. 30. 08. 86, Опубл. 01. 06. 88. 41. Способ сочленения дверной рамы с дверью для создания РЧ – уплотнения. Заявка № 0269205, ЕПВ (ЕР), МКИ4 H05K 9/00, E06B 5/18, 3/12. «The Marconi Company Limited», Заявл. 30. 08. 87, Опубл. 01. 06. 88.

УДК 681.3, УДК 681.142.2, УДК 681.142.4

АНАЛІЗ СУЧАСНИХ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМ В ОБЧИСЛЮВАЛЬНІЙ МЕРЕЖІ ІНТЕРНЕТ

Євген Маєвський
Національний технічний університет України "КПІ"

Анотація: **Розглядаються основні компоненти і надається короткий аналіз сучасних розповсюджених інформаційно-пошукових систем в Інтернет як об'єктів захисту інформаційних технологій, надаються базові правила захисту.**
Summary: **The basic components are considered (examined) and the brief analysis modern popular information systems in the Internet as objects of protection of information technologies is given.**
Ключові слова: **Інформаційно-пошукові системи, пошукові механізми, агенти, пошукові роботи, кроулери, база даних, індексування, сайти, запити, валідність, ранжирування, вагомість.**

Сучасні інформаційно-пошукові системи (ІПС) в обчислювальній мережі Інтернет аналізуються у такій послідовності – пошукові системи чи механізми, пошукові роботи, порівняльний аналіз найбільш поширених ІПС, визначення в компонент та процедур як об'єктів захисту.

Основні протоколи, використовувані в Інтернет (надалі також Мережа), не забезпечені достатніми вбудованими функціями пошуку, не говорячи вже про мільйони серверів, що знаходяться в ній. Протокол HTTP, використовуваний в Інтернет, гарний лише у відношенні навігації, що розглядається тільки як засіб перегляду сторінок, але не їхнього пошуку. Те ж саме відноситься і до протоколу FTP, що навіть більш примітивний, чим HTTP.

Через швидкий зріст інформації, доступної в Мережі, навігаційні методи пошуку інформації швидко досягають межі їхніх функціональних можливостей, не говорячи вже про межу їхньої ефективності. Не вказуючи конкретних цифр, можна сказати, що потрібну інформацію вже не представляється можливим одержати відразу, тому що в Мережі зараз знаходяться мільярди документів і усі вони в розпорядженні користувачів Інтернет, до того ж сьогодні їхня кількість зростає експоненціально. Кількість змін, яким ця інформація піддана, величезна і, саме головне, зміни відбулися за дуже короткий період часу. Основна проблема полягає в тому, що єдиної повної функціональної системи відновлення і занесення подібного обсягу інформації, одночасно доступного всім користувачам Інтернет в усьому світі, ніколи не було.

Для того, щоб структурувати інформацію, накопичену в мережі Інтернет, і забезпечити її користувачів зручними засобами пошуку необхідних їм даних і були створені інформаційно-пошукові системи.

Такі системи принципово складаються з трьох компонентів:

-агента (чи павука, чи кроулера, чи робота), що переміщується Мережею і збирає інформацію;

-бази даних, що містять всю інформацію, яка збирається агентами;

- пошукового механізму, який користувачі Мережі використовують як засіб пошуку і структурування інформації та як інтерфейс для взаємодії з базою даних.

Отже, можна сформулювати перше правило захисту інформаційно-пошукових систем.

Правило ІТ-1. *Основними структурними компонентами захисту в інформаційно-пошукових системах визначаються агенти, бази даних і пошукові механізми. При цьому основними функціями захисту мають бути захист інформаційних технологій (ІТ-продуктів) від загроз цілісності та доступності інформації [2].*

Усі пошукові системи принципово працюють наступним чином.

Пошукові механізми, які іноді називаються і як засоби пошуку і структурування, використовуються для того, щоб допомогти користувачам Інтернет знайти інформацію, у якій вони зацікавлені. Засоби пошуку типу агентів (павуки, кроулери, WEB-роботи) використовуються для збору інформації про документи, що знаходяться в мережі Інтернет. Це спеціальні програми, що ведуть пошук сторінок у Мережі, витягають гіпертекстові посилання на цих сторінках і автоматично індексують інформацію, яку вони знаходять для побудови бази даних.

Кожен пошуковий механізм має власну низку правил, що визначають, як збирати документи. Деякі випливають за кожним посиланням на кожній знайденій сторінці і потім, у свою чергу, досліджують кожне посилання на кожній з нових сторінок, і так далі. Деякі ігнорують посилання, що ведуть до графічних і звукових файлів, файлів мультиплікації; інші ігнорують посилання до ресурсів типу баз даних WAIS; інші працюють за правилом - потрібно переглядати, насамперед, найбільш популярні сторінки.

Агенти – "найінтелектуальніші" з пошукових механізмів. Вони можуть робити більше, ніж просто шукати: вони можуть виконувати навіть транзакції від Вашого імені. Уже зараз вони можуть шукати сайти специфічної тематики і повертати списки сайтів, відсортованих за їх відвідуваністю. Агенти можуть обробляти зміст документів, знаходити й індексувати інші види ресурсів, не тільки сторінки. Вони можуть також бути запрограмовані для витягу інформації з вже існуючих баз даних. Незалежно від інформації, що агенти індексують, вони передають її назад базі даних пошукового механізму.

Як вже зазначалось, загальний пошук інформації в Мережі здійснюють програми, відомі як павуки, кроулери, роботи. Павуки повідомляють про зміст знайденого документа, індексують його і витягають підсумкову інформацію. Також вони переглядають заголовки, деякі посилання і посилають проіндексовану інформацію базі даних пошукового механізму. Кроулери переглядають заголовки і звертають увагу тільки на перше посилання. Пошукові роботи можуть бути запрограмовані так, щоб обслуговувати різні посилання різної глибини вкладеності, виконувати індексацію і навіть перевіряти їх в документі. Внаслідок їхньої природи вони можуть затримуватись в циклах, тому, обслуговуючи посилання, їм потрібні значні ресурси Мережі. Однак, існують методи, призначені для того, щоб заборонити роботам пошук сайтами, власники яких не бажають, щоб вони були проіндексовані.

Агенти витягають і індексують різні види інформації. Деякі, наприклад, індексують кожне окреме слово в документі, що зустрічається, у той час як інші індексують тільки найбільш важливі 100 слів у кожнім документі, індексують розмір документу і число слів у ньому, назву, заголовки і підзаголовки і так далі. Вид побудованого індексу визначає, який пошук може бути зроблено пошуковим механізмом і як отримана інформація буде інтерпретована.

Агенти можуть також переміщатися Мережею і знаходити інформацію, після чого поміщати її в базу даних пошукового механізму. Адміністратори пошукових систем можуть визначити, які сайти чи типи сайтів агенти повинні відвідати і проіндексувати. Проіндексована інформація відсилається базі даних пошукового механізму так само, як було описано вище.

Користувачі Мережі можуть поміщати інформацію прямо в індекс, заповнюючи особливу форму для того розділу, у який вони хотіли б помістити свою інформацію. Ця інформація передається базі даних.

Коли хто-небудь хоче знайти інформацію, доступну в Інтернет, він відвідує сторінку пошукової системи і заповнює форму, що деталізує інформацію, яка йому необхідна. Тут можуть використовуватися ключові слова, дати й інші критерії. Критерії у формі пошуку мають відповідати критеріям, використовуваним агентами при індексації інформації, що вони знайшли при переміщенні Мережею.

База даних відшукує предмет запиту, заснований на інформації, зазначеній у заповненій формі, і виводить відповідні документи, підготовлені базою даних. Щоб визначити порядок, у якому список документів буде показаний, база даних застосовує алгоритм ранжирування. В ідеальному випадку, документи, найбільш валідні користувальницькому запиту, будуть поміщені першими в списку. Різні пошукові системи використовують різні алгоритми ранжирування, однак основні принципи визначення валідності наступні:

-кількість слів запиту в текстовому змісті документу (тобто, в html-коді);

-теги, у яких ці слова розташовуються;

-місце розташування шуканих слів у документі;

-питома вага слів, щодо яких визначається валідність, у загальній кількості слів документу;

-час – як довго сторінка знаходиться в базі даних пошукового серверу. Спочатку здається, що це досить безглуздий принцип. Але в Інтернеті існує багато сайтів, що живуть максимум місяць! Якщо ж сайт існує досить довго, то це означає, що власник дуже досвідчений у даній темі і користувачу більше підходить сайт, що пару років віщає світу про правила поведінки за столом, чим той, що з'явився тиждень тому на ту ж тему;

-індекс цитатності – як багато посилань на дану сторінку з інших сторінок, зареєстрованих у базі даних під час пошуку.

Ці принципи застосовуються всіма пошуковими системами, а представлені нижче – деякими, але досить відомими системами (зокрема, AltaVista, HotBot).

База даних виводить ранжирований подібним чином список документів з HTML і повертає його користувачу, що зробив запит. Різні пошукові механізми також вибирають різні способи показу отриманого списку – деякі показують тільки посилання; інші виводять посилання с першими декількома пропозиціями, що містяться в документі чи у заголовку документа разом з посиланням.

Коли ви звертаєтесь до одного з документів, що вас цікавить, то цей документ запитується у того сервера, на якому він знаходиться. Але в усіх випадках, коли визначається певне посилання, тобто запит обслуговування його пошуковою системою, нас, насамперед, цікавить одне – до послуг якого серверу слід звертатись і яку пошукову систему слід обрати.

Далі надається порівняльний аналіз роботи найбільш поширених пошукових систем.

1. Lycos. Lycos використовує наступний *механізм індексації*:

-індекси слів в <title> заголовку мають вищий пріоритет;

-індексація слова на початку сторінки;

-індексація слів в посиланнях;

Якщо в базі даних щодо індексу є сайти, посилання з яких вказує на індексований документ, то *валідність* цього документу зростає.

Як і більшість систем, Lycos дає можливість застосовувати простий запит і більш витончений метод пошуку. У простому запиті як пошуковий критерій вводиться пропозиція природною мовою, після чого Lycos робить нормалізацію запиту, видаляючи з нього так звані stop-слова, і тільки після цього приступає до його виконання. Майже відразу видається інформація про кількість документів на кожне слово, а пізніше і список посилань на формально релевантні документи. У списку проти кожного документа вказується його міра близькості запиту, кількість слів із запиту, що потрапили в документ, і оцінена міра близькості, що може бути більшою чи меншою за формально обчислену. Поки не можна вводити логічні оператори в рядку разом з термінами, але використовувати логіку через систему меню Lycos дозволяє. Така можливість застосовується для побудови розширеної форми запиту, призначеної для спокушених користувачів, що вже навчилися працювати з цим механізмом. Таким чином, видно, що Lycos відноситься до системи з мовою запитів типу "Like this", але намічається його розширення і на інші способи організації пошукових розпоряджень.

2. AltaVista. Індексування в цій системі здійснюється за допомогою роботу. При цьому робот має наступні пріоритети:

-слова, що містяться в тегі <title>, мають вищий пріоритет;

-ключові фрази в <Meta> тегах;

-ключові фрази, що знаходяться на початку сторінки;

-ключові фрази в ALT – посиланнях;

-ключові фрази за кількістю входжень\присутності слів\фраз.

Якщо тегів на сторінці пошуку немає, то використовуються перші 30 слів, що індексуються і показуються замість опису (tag description).

Найбільш цікава можливість AltaVista – це розширений пошук. Тут варто відразу зауважити, що, на відміну від багатьох інших систем, AltaVista підтримує одномістний оператор NOT і оператор NEAR, що реалізує можливість контекстного пошуку, коли терміни можуть розташовуватися поруч у тексті документа.

AltaVista дозволяє пошук ключовими фразами, при цьому вона має досить великий фразеологічний словник. Крім усього іншого, під час пошуку в AltaVista можна задати ім'я полю, де має зустрітися слово: гіпертекстове посилання, applet, назва образу, заголовок і ряд інших полів. На жаль, докладно процедура ранжирування в документації на систему не описана, але видно, що ранжирування застосовується як при простому пошуку, так і при розширеному запиті. Реально цю систему можна віднести до системи з розширеним булєвим пошуком.

3. **Yahoo.** Дана система з'явилася в Мережі однією з перших і сьогодні Yahoo співробітничає з багатьма виробниками засобів інформаційного пошуку, а на різних її серверах використовується різне програмне забезпечення. Мова Yahoo досить проста: усі слова варто вводити через пробіл, вони з'єднуються зв'язуванням AND або OR. При видачі не вказується ступінь відповідності документа запиту, а тільки підкреслюються слова з запиту, що зустрілися в документі. При цьому не виробляється нормалізація лексики і не проводиться аналіз на "загальні" слова. Гарні результати пошуку виходять тільки тоді, коли користувач знає, що в базі даних Yahoo інформація є напевно. *Ранжирування* виробляється за числом термінів запиту в документі. Yahoo відноситься до класу простих традиційних систем з обмеженими можливостями пошуку.

4. **OpenText.** Інформаційна система OpenText є найбільш комерціалізованим інформаційним продуктом у Мережі. Всі описи більше схожі на рекламу, чим на інформативний посібник з роботи. Система дозволяє провести пошук з використанням логічних конекторів, однак розмір запиту обмежений трьома термінами чи фразами. У даному випадку мова йде про розширений пошук. При видачі результатів повідомляється ступінь відповідності документа запиту і розмір документу. Система дозволяє також поліпшити результати пошуку в стилі традиційного булєвого пошуку. OpenText можна було б віднести до розряду традиційних інформаційно-пошукових систем, якби не *механізм ранжирування*.

5. Infoseek. У цій системі індекс створює робот, але він *індексує* не весь сайт, а тільки зазначену сторінку. При цьому робот має такі пріоритети:

-слова в заголовку <title> мають найвищий пріоритет;

-слова в тегі keywords, description і частота входжень\повторень у самому тексті мають менший пріоритет;

-при повторенні однакових слів поруч – викидає з індексу.

Infoseek допускає до 1024 символи для тега keywords, 200 символів для тега description. Якщо теги не використовувалися, індексує перші 200 слів на сторінці і використовує як опис. Система Infoseek має досить розвинену інформаційно-пошукову мову, що дозволяє не просто вказувати, які терміни можуть зустрічатися в документах, але і своєрідно зважувати їх. Досягається це за допомогою спеціальних знаків "+" – термін зобов'язаний бути в документі, і "–" – термін обов‘язково має бути відсутнім у документі.

Крім того, Infoseek дозволяє реалізувати те, що називається контекстним пошуком. Це значить, що використовуючи спеціальну форму запиту, можна зажадати послідовної спільної зустрічальності слів. Також можна вказати, що деякі слова можуть одночасно зустрічатися не тільки в одному документі, а навіть в окремому параграфі чи заголовку. Є можливість указівки на ключові фрази, що представляють собою єдине ціле, аж до порядку слів. *Ранжирування* при видачі здійснюється за кількістю термінів запиту в документі, за кількістю фраз запиту, за винятком загальних слів. Усі ці фактори використовуються як вкладені процедури.

Як коротке резюме, можна сказати, що Infoseek відноситься до традиційних систем з елементом *зважування* термінів під час пошуку.

6. WAIS. WAIS є однією з найбільш витончених пошукових систем Internet. У ній не реалізовано лише пошук щодо нечітких множин і ймовірний пошук. На відміну від багатьох пошукових машин, система дозволяє будувати не тільки вкладені булєві запити, рахувати формальну *валідність* за різними мірками близькості, зважувати терміни запиту і документу, але і здійснювати корекцію запиту щодо валідності. Система також дозволяє використовувати усікання термінів, розбивку документів на поля і ведення розподілених індексів. Не випадково саме ця система була обрана як основний пошуковий механізм для реалізації енциклопедії "Британіка" на Internet.

На основі проведеного аналізу роботи популярних інформаційно-пошукових систем можна сформулювати друге правило їх захисту.

Правило ІТ-2. *Основними функціональними компонентами захисту в інформаційно-пошукових системах Інтернет доцільно визначати механізми індексування, валідності, вагомості та ранжирування компонентів запитів користувача. При цьому основними функціями захисту мають бути захист інформаційних технологій (ІТ-продуктів) від загроз цілісності та доступності інформації.*

Висновки

1. Визначені у статті базові правила захисту інформаційних технологій щодо інформаційно-пошукових систем у мережі *Інтернет* у постановчому плані консалтингово орієнтують можливі напрями забезпечення національної безпеки України в інформаційній та науково-технологічній сферах.

2. Проблема забезпечення національної безпеки України в міжнародному інформаційному просторі визначається пріоритетною та однією із семи її сфер. Визначені у статті правила захисту інформаційних технологій щодо пошукових систем в Інтернет можуть бути використані для подальших досліджень та шляхів рішення цієї проблеми і у мережах Інтранет, Екстранет.

*Література: 1. Дєвид Стенг, Сильвия Мунг. Секреты безопасности сетей. *ICE "Диалектика", "Информейшн Компьютер Энтерпрайз". Киев, 1996. 2. В. В. Шорошев, А. Э. Ильницкий, И. Л. Близнюк, И. О. Бакаев, Н. Г .Панько. Проблемные вопросы стратегии безопасности корпоративных компьютерных систем. Бизнес и безопасность. № 5, 2000. 3. С. Золотов. Протоколы Internet. – СПб.; BHV – Санкт-Петербург, 1998. 4. http://www.citforum.ru/internet/search/searchsystems.shtml. Поисковые системы в сети Интернет. © В. Тихонов, Ноябрь 2000 atomzone.hypermart.net*

УДК 681.3, УДК 681.142.2, УДК 681.142.4

МЕТОДИЧЕСКИЕ ОСНОВЫ ЭФФЕКТИВНОГО ПОИСКА ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Вячеслав Шорошев, Евгений Маевский*

*НИИ НАВД Украины, *Национальный технический университет Украины "КПИ"*

Аннотация: **Рассматриваются автоматический анализ текста на основе первого и второго законов Зипфа, весовые коэффициенты слов-терминов, матричное и векторное представление базы данных поисковых систем как объектов защиты информационных технологий.**

*Summary: A***re considered the automatic analysis of the text on the basis of first and second laws of Zipf, weight factors of words - terms, matrix and vector representation of a database of search systems.**

Ключевые слова: **Законы Зипфа, ранг, слова-термины, матричное представление данных, пространсвенно-векторное представление данных, релевантность, инверсная частота.**

Методы эффективного и надежного поиска информации в Интернет имеют специфические особенности. Для выяснения сущности этого процесса необходимо знать, как функционирует информационно-поисковая система (ИПС), и решает задачи обеспечения защиты информации [1–3]. Изложим некоторые методические основы эффективного поиска информации в Интернет на примере одной из базовых методик [4].

Целью статьи является определение базовых, рекомендательных правил или направлений возможных решений задач поиска для обеспечения информационной безопасности.

Для этого рассмотрим ряд аспектов информационного поиска, в том числе на основании универсальных законов Зипфа, которые используем для защиты информационных технологий поисковых систем.

Автоматический анализ текстов

Оказывается, что все созданные человеком тексты построены по единым правилам. Никому не удается обойти их. Какой бы язык не использовался, кто бы ни писал, даже классик графоман, внутренняя структура текста останется неизменной. Она описывается так называемыми законами Зипфа (George. K. Zipf). Зипф предположил, что естественная человеческая лень ведет к тому, что слова с большим количеством букв встречаются в тексте реже коротких слов. Базируясь на этом постулате, Зипф вывел два универсальных закона.

Первый закон Зипфа – "ранг – частота"

Выберем слово и сосчитаем, сколько раз оно встречается в тексте. Эта величина называется частотой вхождения слова. Определим частоту вхождения каждого слова в тексте. Некоторые слова будут иметь одинаковую частоту. Сгруппируем их, взяв только одно значение из каждой группы. Расположим частоты по мере их убывания и пронумеруем. Порядковый номер частоты определяется как *ранг частоты* (рис. 1). Так, если наиболее часто встречаются определенные слова, то они будут иметь ранг 1, а те, которые появляются за ними, – ранг 2 и т. д.

Выберем наугад страницу и определим вероятность встречи слова, на которое выпал выбор. Вероятность будет равняться отношению частоты вхождения этого слова к общему числу слов в тексте

Вероятность = Частота вхождения слова / Число слов.

Зипф выявил интересную закономерность. Оказывается, что если помножить вероятность выявления слова в тексте на ранг частоты, то величина, которая получается (это "С" в формуле) приблизительно постоянная.

С = (Частота вхождения слова х Ранг частоты) / Число слов

Это функция типа $y = k/x$ и ее график – гипербола. Итак, по первому закону Зипфа, если наиболее распространенное слово встречается в тексте, например, 100 раз, то следующее по частоте слово едва ли встретится 99 раз. Частота вхождения второго по популярности слова с высокой долей вероятности окажется на уровне 50.