УДК 51.001.57+681.142.2

Ю.О. Сєров
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ТЕХНОЛОГІЇ ПОШУКУ ТА ВИДОБУВАННЯ ДАНИХ У WWW (АНАЛІЗ ПРОБЛЕМИ)

This paper considers information in Web finding and Web mining technologies analyzing. In this paper information finding methods and problems of these methods are analyzed. Web mining technologies are described and web mining categorization is made. Perspective web mining technologies are studied especially structured query languages applied on the Web.

Розглянуто актуальну проблему пошуку інформації у Веб та досліджень технології web mining. Аналізуються способи пошуку інформації та проблеми, які виникають. Описуються технології web mining, поділ web mining на категорії. Досліджуються перспективні напрямки технології web mining, зокрема технології структурованих мов запитів до Веб.

ПОСТАНОВКА ПРОБЛЕМИ В ЗАГАЛЬНОМУ ВИГЛЯДІ

World Wide Web (WWW) містить надзвичайно багато інформації (проіндексований Веб містить приблизно 3 мільярди сторінок – приблизно 25–50 терабайт інформації, непроіндексований – у 400–550 разів більше) і її кількість зростає з кожним днем (щодня з’являється близько 7 мільйонів нових сторінок).

Тому під час роботи з Веб користувачі, що шукають інформацію, стикаються з такими проблемами:

1. Знаходження релевантної інформації. Користувачі Інтернет переглядають Веб-сайти або користуються пошуковими сервісами для знаходження необхідної інформації. При використанні пошукових сервісів користувач задає запит за ключовими словами, і відповіддю на цей запит є впорядкований список сторінок, котрі містять задані ключові слова. Однак сучасні пошукові засоби мають ряд недоліків: низька релевантність багатьох знайдених ресурсів, незнаходження потрібної інформації через непроіндексованість потрібних сторінок.

2. Виведення знань з інформації, яка міститься у Веб. Взагалі кажучи, ця проблема тісно пов’язана з попередньою. Вважається, що вже існує масив даних з Веб і потрібно вивести знання, які будуть корисні.

3. Персоналізація інформації. Проблема пов’язана з типом і презентацією інформації, оскільки користувачі відрізняються тим, яка інформація їм потрібна і як вони хочуть її бачити. Тому для людей, котрі хочуть досягнути своєї мети, необхідно знати: яку інформацію шукають і як її краще представити кожному конкретному користувачу.

4. Знання про споживачів/користувачів. Ця проблема тісно пов’язана з попередньою, яка полягає у вивченні того, що роблять клієнти і чого вони хочуть. Ця проблема складається з менших: надання інформації конкретному користувачу, дизайн Веб-сайту, проблеми менеджменту і маркетингу.

Сьогодні одним із перспективних напрямків розвитку технологій обробки інформації у Веб є методи Web mining [1].

Методи Web mining можуть бути використані для вирішення цих проблем, хоча вони і не є єдиним засобом. Одночасно з Web mining можуть бути використані такі методики, як: бази даних, пошук інформації (informational retrieval), обробка природної мови (natural language processing).

1. АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

1.1. Організація інформації і доступ до неї

1.1.1. Пошук

Пошук – найпростіший спосіб доступу до текстових даних. Сучасні пошукові машини мають такі можливості:

- індексування тексту;
- пошук за ключовими словами;
- морфологічний пошук – пошук за словоформами;
- логічна мова запитів, яка дозволяє задавати умови спільного входження (невходження) ключових слів у документ;
- рангування документів відповідно до ключового запиту.

Однак, при сьогоднішніх швидкостях зростання інформації в Інтернет цих можливостей вже не вистачає. Тому сучасні пошукові машини оснащуються додатковими засобами пошуку.

Одним з цих засобів є функція “знайти схоже”. Ця функція дає можливість уточнити запит – виділити один чи кілька знайдених документів і знайти документи, схожі на виділені. Цією функцією володіють практично усі сучасні пошукові машини: Google, Alta Vista, HotBot, Yandex [2].

1.1.2. Пошук за вибіркою

Функція “пошук за вибіркою” дає можливість задавати додаткові обмеження пошуку за множиною документів, які вже знайдені за простішим запитом. Використання цієї функції необхідне, якщо простіший запит дав дуже багато документів і потрібно “звузити” запит і тим самим зменшити кількість знайдених документів.

1.1.3. Запит звичайною мовою

На відміну від логічних запитів запит природною мовою дозволяє користувачу задати запитання так, як він задав би його людині. Наприклад: “Чи є життя на Марсі?”. Ця функція реалізується шляхом відкидання “шумових слів” і виразів (“чи”, “є”, “на”), підстановкою синонімів, пошуком за різними словоформами. Дуже багато пошукових машин заявляють про наявність такої функції, але дуже часто ця функція не працює.

1.1.4. Тезауруси

Тезауруси (словники) призначені для розширення запиту і містять синоніми, антоніми, однокореневі слова тощо. Однак більшість сучасних пошукових машин тезауруса не мають. Мабуть, це зумовлено дороговизною і складністю лексикографічної роботи.

1.2. Каталоги

Ще одним поширеним способом організації інформації в Інтернет є створення каталогів (рубрикаторів і класифікаторів). Створюються вони просто: задається дерево

категорій, яке потім заповнюється посиланнями на документи з коротким описом змісту документу.

Проблемою створення каталогів є те, що всю роботу з класифікації і написання анотації необхідно виконувати вручну, через те, що задачу класифікації ресурсу дуже важко автоматизувати. Наприклад, над створенням таких каталогів, як ODP (http://dmoz.org/) працюють десятки тисяч людей, які є спеціалістами в тій чи іншій царині. Хоча зараз вже з'являються засоби автоматизації цього процесу.

Каталоги зручні тим, що там міститься "якісніша" інформація, але для ефективного пошуку в каталозі треба знати принцип структуризації, який дуже часто відомий лише творцям цього каталогу. Вирішенням цієї проблеми є приєднання до каталогу засобів пошуку або каталогу до пошукової машини.

1.3. Анотування

Існуючі каталоги дають можливість анотувати занесені в них посилання. З точки зору пошуку інформації анотування є дуже зручним, оскільки набагато зменшує час пошуку інформації. Проблемою анотування є те, що анотації створюються вручну. Кількість інформації зростає експоненціально, тому неможливо встигати анотувати всю інформацію, що з'являється. Тому виникає необхідність автоматизації каталожної роботи, зокрема в створенні анотацій.

2. ФОРМУВАННЯ ЦІЛЕЙ

У статті розглядаються такі питання:

- Аналіз сучасного стану засобів пошуку в Веб.
- Аналіз методів видобування інформації з Веб та сучасних досліджень у цьому напрямку.
- Розгляд недоліків існуючих технологій.
- Огляд категорій Web mining (Web content mining, Web structure mining, Web usage mining).
- Перспективи розвитку технологій Web mining (видобування фактів, WebSQL).

3. ОСНОВНИЙ МАТЕРІАЛ

3.1. Web mining

Web mining – це використання методів data mining для автоматичного знаходження і видобування (extraction) інформації з Веб-документів і сервісів. У цьому напрямку проводиться дуже багато досліджень через неймовірне зростання кількості інформації, яка з'являється в Веб і великою зацікавленістю в електронній комерції (e-commerce).

Web mining складається з таких підзадач:

1. Знаходження ресурсів – знаходження відповідних Веб-документів. Знаходження ресурсів – це процес знаходження і вибору даних з текстових джерел, які містяться у Веб: таких як електронні новини, групи новин, текстове наповнення HTML з відкиданням тегів, а також ручна вибірка Веб-ресурсів. До ресурсів також належать он-лайнові текстові ресурси, зроблені для дослідників, текстові бази даних тощо.

2. Вибірка(Selection) інформації і попередня обробка – автоматична вибірка і попередня обробка інформації із знайдених Веб-ресурсів.

3. Узагальнення(Generalization) – автоматичне знаходження загальних закономірностей (pattern) як на окремих Веб-сайтах, так і на множині сайтів.

4. Аналіз – перевірка і/або інтерпретація знайдених закономірностей.

Web mining тісно пов'язаний з машинним навчанням і аналізом даних. Також Web mining часто асоціюється з пошуком інформації (Information Retrieval) та видобуванням інформації (Information Extraction), хоча насправді це не те саме.

3.1.1. Web mining і пошук інформації (Information Retrieval)

Пошук інформації (Information Retrieval – IR) – автоматичне знаходження усіх релевантних документів і водночас мінімізація нерелевантних документів серед знайдених, а також рангування знайдених документів за мірою релевантності. Первинна мета IR – індексування текстів і пошук важливих документів. Сучасні дослідження також відносять до IR моделювання, класифікацію документів, інтерфейси користувача, візуалізацію даних, фільтрування тощо. Якщо вважати, що Web mining – класифікація Веб-документів з подальшою індексацією, то тоді Web mining є частиною процесу IR. У будь-якому випадку задачі індексування використовують методи data mining.

3.1.2. Web mining і видобування інформації (Information Extraction)

Видобування інформації (Information Extraction – IE) полягає в обробці колекцій документів і представленні інформації, яку вони містять, у формі, зручній для роботи і аналізування. На відміну від IR, метою якого є знаходження релевантних документів, метою IE є знаходження релевантних даних у документах. IE полягає в аналізі структури і представленні документів, а IR – в знаходженні множини невпорядкованих даних.

3.1.3. Web Mining і застосування машинного навчання у Веб (Machine Learning Applied on the Web)

Web mining – це не те саме, що навчання з Веб чи методи машинного навчання, застосовані до Веб.

З одного боку, існують застосування (application) з машинним навчанням, застосованим до Веб, які не є зразками Web mining. Прикладом цього є методологія машинного навчання, котра використовується для пошуку у Веб за заданою тематикою, яка робить акцент на знаходженні наступних напрямків пошуку.

З іншого боку, Web mining використовує не лише методи машинного навчання. Прикладом цього є специфічні алгоритми видобування інформації з авторитетних сайтів (authority) чи вказівних сайтів (hub) і дослідження схеми Веб (Web scema).

Однак, незважаючи на це, Web mining і машинне навчання дуже близькі області досліджень і машинне навчання застосовується у Web mining. Наприклад, нещодавні дослідження [8] показують, що застосування машинного навчання може покращити процес класифікації текстів порівняно зі звичайними IR технологіями.

Отже, Web mining перетинається із застосуванням машинного навчання у Веб.

3.2. Категорії Web mining

Web mining поділяється на 3 категорії, відповідно до частин Веб, які можна досліджувати: дослідження вмісту Веб (Web content mining), дослідження структури Веб (Web structure mining), дослідження поведінки користувачів (Web usage mining).

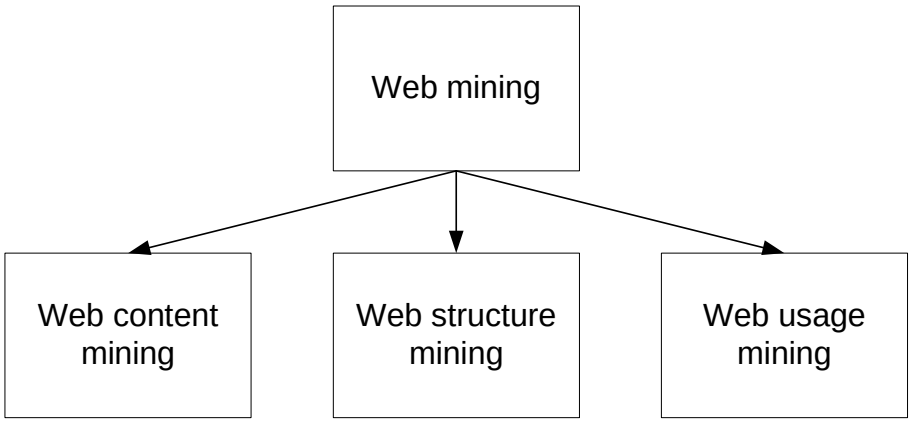

Рис. 1. Категорії web mining

3.2.1. Дослідження вмісту Веб (Web content mining)

3.2.1.1. Загальний опис Web content mining

Web content mining займається знаходженням потрібної інформації у Веб-документах і даних, які містяться у Веб.

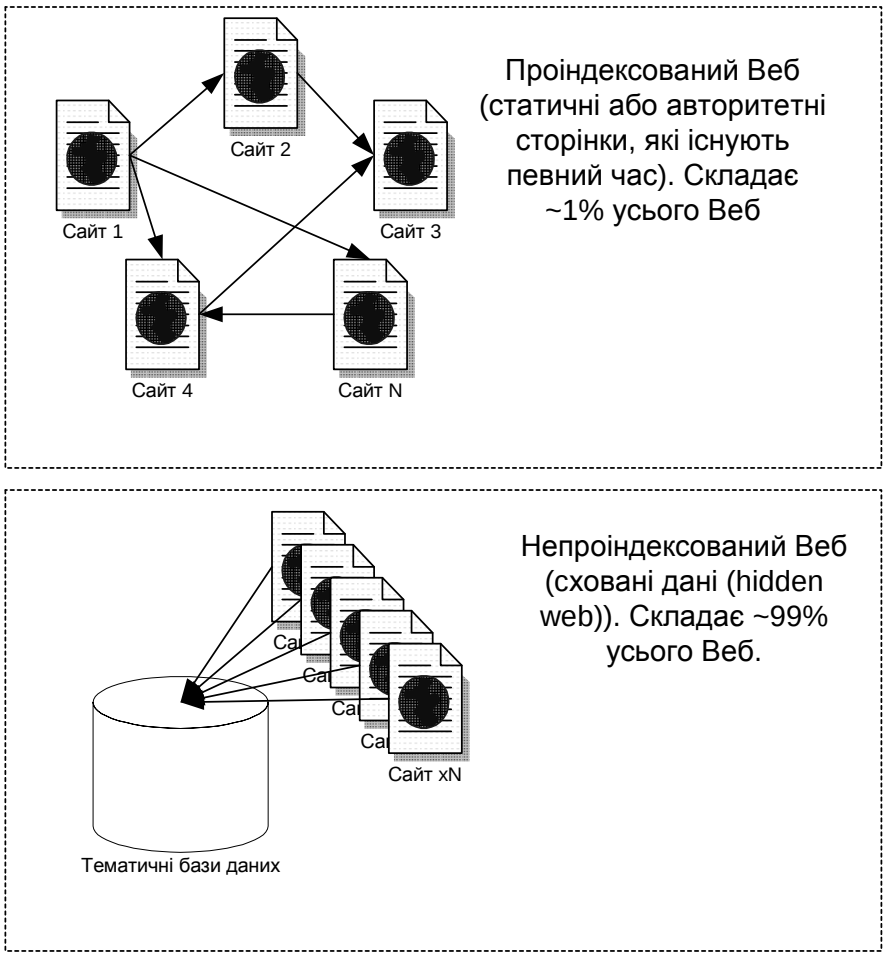

Рис. 2. Проблема схованих даних у WWW

Кількість інформації у Веб постійно зростає. Це відбувається за рахунок багатьох чинників:

- Раніше Інтернет складався з різних типів сервісів та джерел даних, таких як Gopher, FTP і Usenet, а тепер ці джерела переміщаються у Веб.
- Електронні бібліотеки, доступні у Веб.

Багато компаній переносять свій бізнес у Веб і, як наслідок, роблять свої бази даних доступними у Веб. Тобто працівники, партнери чи навіть клієнти можуть мати доступ до них через Веб інтерфейси.

Крім того, у Веб багато “схованих” даних (hidden web), які не можуть бути проіндексовані. Це дані, котрі містяться у різноманітних тематичних базах даних і генеруються динамічно як результат запиту до СУБД (наприклад, бази даних географічних карт, прогнозів погоди, телефонні довідники тощо) або є приватними. Схованими є також дані на авторитетних сайтах, які ще не були проіндексовані (форуми, чати, сайти інформаційних агентств). Об’єм схованих даних становить приблизно 7500 терабайт інформації.

Тому основне завдання Web content mining – видобування знань з усіх даних, які містяться у Веб.

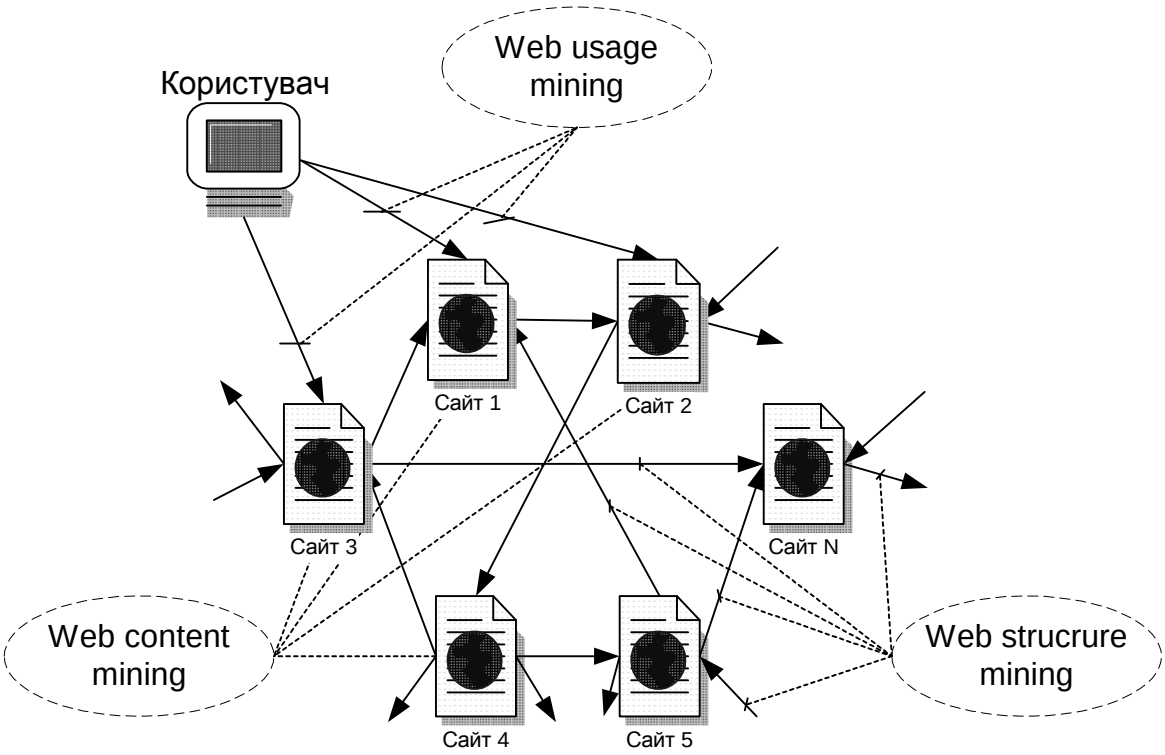

Рис. 3. Поділ web mining за частинами Веб, які досліджуються

3.2.1.2. Пошук особливих контекстних явищ

Сучасні пошукові машини не є інтелектуальними, що може стати проблемою при пошуку певної інформації. Наприклад: потрібно знайти всі документи, в яких згадується дата 7 червня 2003. Більшість пошукових машин з цією задачею не впораються.

По-перше, 7 червня 2003 – це три слова: “7”, “червня”, “2003”. По-друге, ця дата може бути записана кількома способами: 07.06.2003, сьоме червня 2003 тощо.

Такі ситуації трапляються дуже часто, а для користувача дуже часто важливими є саме такі контекстні явища: дати, суми, номери телефонів, імена людей (прізвище, ім’я, по батькові), назви фірм, товарів тощо.

Ці недоліки пошукових систем вирішуються шляхом написання спеціальних “розпізнавачів” для кожного типу контекстних явищ.

На жаль, жодна сучасна пошукова машина в Інтернет не обладнана таким розпізнавачем.

3.2.1.3. Фактографічні запити

Важливою і необхідною рисою сучасної пошукової машини є можливість відповідати на фактографічні запити. Наприклад: “Де придбати дешевий вживаний комп’ютер”. Особливо важливою вона є для електронних енциклопедій, служб технічної підтримки, мережних бібліотек.

Ця задача є дуже складною, оскільки на відміну від звичайного пошуку входження ключових слів у документи, для того, щоб система знайшла правильну відповідь на запитання “Хто був президентом України в минулому столітті” потрібно, щоб вона правильно замінила слова “в минулому столітті” на потрібний часовий інтервал, а також передивилася багато документів, об’єднала їх у єдину відповідь.

Очевидно, що ця задача є надзвичайно складною, і сьогодні не існує програм, які б дозволяли задавати фактографічні запити до текстових баз даних.

3.2.1.4. Ланцюжки фактів

Ще однією важливою задачею, яку вирішує Web mining, є побудова “ланцюжків фактів”.

“Ланцюжок фактів” – збір з множини документів окремих даних, між якими існує певний зв’язок, з яких ми отримуємо нові знання.

Наприклад:

“Громадянин, який звернувся до нашого банку за кредитом у 5 млн. грн. від ЗАТ “X”, у 1995 був одружений з власницею ЗАТ “Y”.

“ЗАТ “Y” не повернуло кредит у розмірі 2 млн. грн.”

Звідси ми можемо зробити висновок: дуже велика ймовірність, що наш клієнт – шахрай, тому не варто давати йому кредит.

Основною проблемою у вирішенні цієї задачі є розпізнавання об’єкта дослідження і зведення до канонічної форми запису.

3.2.2. Дослідження структури Веб (Web structure mining)

Завданням Web structure mining є відтворення моделі Веб, яка відображає топологію гіперспосилань з описом цих посилань чи без опису. Ця модель використовується для класифікації Веб-сторінок і дає корисну інформацію про взаємовідношення між Веб-сайтами та їх подібність між собою.

Web structure mining також використовується для знаходження авторитетних Веб-сайтів (authority) чи вказівних сайтів (hub) (вказують на авторитетні сайти).

Цей напрям досліджень тісно пов'язаний з аналізом цитат та соціальних мереж (social networks and citations analysis). Аналіз соціальних мереж застосовується для знаходження різних типів сайтів (тематичні, вказівні) за допомогою вхідних і вихідних посилань. Web structure mining застосовує аналіз соціальних мереж до структури гіперпосилань у Веб для знаходження певних зв'язків. Наприклад, у своєму дослідженні Kautz [4] застосовує аналіз до об'єкта "людина" для моделювання мережі дослідників штучного інтелекту. Для цього використовуються дані про людей, знайдені на домашніх сторінках, форумах, гіперпосилання з домашніх сторінок, дані про обмін інформацією, знайдені у архівах тощо. Такі дослідження є комбінацією Web content і Web structure mining.

3.2.3. Дослідження поведінки користувачів (Web usage mining)

Завданням Web usage mining є дослідження даних, згенерованих під час сесій користувачів Інтернет та їхньої поведінки. На відміну від Web content mining і Web usage mining, які використовують первинні дані Веб, Web usage mining досліджує вторинні дані про дії користувачів в Інтернеті. Web usage mining використовує дані про доступ до Веб-серверів (логи), Проксі-серверів, профілі користувача, дані реєстрацій, дані про сесії користувача, закладки у брозері, клацання миші тощо.

За джерелами даних, які використовуються, Web usage mining можна класифікувати:

Веб-клієнти;

Проксі-сервери;

Сервери;

Web usage mining може відбуватися двома шляхами: перетворення даних з Веб-сервера у реляційні таблиці перед тим, як застосовувати data mining або пряме застосування спеціальних методологій попередньої обробки.

Застосування Web usage mining поділяється на дві категорії:

Вивчення профілю користувача чи моделювання користувача;

Вивчення навігаційних зразків користувача.

Web usage mining є корисним для всіх сторін. Користувачі зацікавлені в отриманні потрібної інформації у вигляді, який їм подобається. Провайдери інформації (Веб-майстри) зацікавлені у тому, щоб їх інформація була ефективною і популярною. А для цього потрібно знати, що подобається користувачам, щоб персоналізувати сайт (адаптувати дизайн сайту згідно з побажаннями користувача).

Слід зауважити, що межа між трьома видами Web mining є досить розмитою. Web content mining може використовувати текст, посилання і навіть профілі користувачів. Web structure mining, крім структури посилань, може використовувати інформацію про посилання. Крім того, можна використовувати дані сервера: відслідковувати посилання, які зацікавили користувача, протягом сесії.

3.3. Перспективи

3.3.1. Вибірка на визначену тему

Сьогодні засоби пошуку і стиснення інформації вже настільки розвинулися, що незабаром з'являться системи, які б готували вибірку на визначену тему за заданою областю даних (базою даних чи Інтернет). Технічно нескладно реалізувати це з використанням "розумної" пошукової машини та існуючих засобів семантичного стиснення інформації і знаходження смислових дублів.

3.3.2. Видобування фактів (вікна фактів)

Метод збору інформації полягає в тому, що з документів видобуваються лише безперечні факти, часто дуже прості і нецікаві. Наприклад, з речення "Прем'єр-міністр України "Y" ствердив, що у 20xx році валовий прибуток зріс на 5 відсотків" можна виділити лише один безсумнівний факт: "У 20xx році прем'єр-міністром України був "Y"".

Виявилося, що співставлення таких простих, атомарних фактів може дати несподівані нові знання. Наприклад, за газетними публікаціями можна дізнатися усю біографію "Y".

Можна припустити, що сучасні пошукові системи перейдуть від простої індексації слів у документах Інтернет до збирання фактів. Технічно це не дуже важко, а атомарних фактів в Інтернет – безліч.

Факти, які збираються таким чином, мають дуже просту структуру, їх легко перетворити в знання і проводити за ними логічне виведення.

3.3.3. WebSQL

Сьогодні технології баз даних є надзвичайно потужним і гнучким засобом для створення запитів до добре структурованих даних [5]. Були зроблені спроби застосувати ці ж технології баз до Веб. Але на цьому шляху є ряд перепон:

- Під час виконання запиту неможливо пронумерувати усі документи у Веб, оскільки вони утворюють невизначену множину сторінок. А навіть, якби це було можливо, це було б практично нереалізовно, через величезну кількість інформації (7500 терабайт), яка міститься у Веб.
- Інформація, яка міститься у Веб, є слабоструктурованою. Мови запитів, розроблені для добре структурованих даних, погано застосовні для створення запитів у Веб.
- Документи HTML дуже часто містять помилки, що спричинює труднощі при видобуванні структури документу.
- Дані зберігаються у файлах різних типів: текстові, графічні, звукові, що ускладнює визначення того, чи задовольняє файл обмеження чи ні.

WebSQL – мова запитів у Веб, спроба застосувати технології реляційних баз даних до Веб. Моделюючи Веб як реляційну базу даних з двома вуртуальними відношеннями Документи і Посилання, можна створювати SQL-подібні запити з обмеженнями на локальність, структуру, тип документа, дату модифікації і зміст. WebSQL – комбінування структурних і змістовних запитів з використанням структури і топології Веб. Ця мова не є

замінником індексних серверів, оскільки інтерфейс користувача є надто складним. Мета використання WebSQL – полегшення розробки застосувань, призначених для селективного індексування, автоматичного створення посилань тощо.

У WebSQL Веб розглядається як віртуальний граф, у якому документи є вершинами, з'єднані посиланнями. Тоді документ у Веб розглядається як кортеж віртуального відношення Документ ([*url, title, text, type, length, modif*]). Url (uniform resource locator) – уніфікований ідентифікатор інформаційного ресурсу – ідентифікує об'єкт, заголовок (*title*) і текст (*text)* містяться в HTML документі, дані про тип (*type*), довжину (*length*) та останню модифікацію (*modif*) містяться на сервері.

Існують також аналогічні підходи, які називаються W3QS, W3QL та інші.

ВИСНОВКИ

Веб містить величезну кількість інформації, об'єм якої зростає з кожним днем. Сучасні засоби пошуку інформації у Веб (пошукові машини, каталоги, тощо) не забезпечують потреб користувачів у знаходженні релевантної інформації.

Web mining – data mining, застосовний на множині даних Веб. Ця технологія покликана розв'язати проблеми знаходження релевантної інформації, виведення знань з інформації, яка міститься у Веб, персоналізація інформації.

Web mining поділяється на 3 категорії: Web content mining, Web structure mining, Web usage mining. Завданням Web content mining є знаходження потрібної інформації у Веб-документах і даних, які містяться у Веб; завданням Web structure mining – відтворення моделі Веб, яка відображає топологію гіперспосилань, Web usage mining – дослідження даних, згенерованих під час сесій користувачів Інтернет.

За приблизними оцінками, 95 відсотків даних, котрі містяться у Веб, є схованими (hidden). Як правило, це дані, котрі зберігаються у різноманітних тематичних базах даних (телефонні довідники, довідники прогнозу погоди тощо). Крім того, схованими є дані, які містяться на авторитетних сайтах, але їх ще не встигли проіндексувати пошукові машини. Тому одним з пріоритетних напрямків досліджень є видобування схованих даних.

Сьогодні одним з найперспективніших напрямків розвитку Web mining є розробка мови запитів до Веб, аналогічної до мови запитів SQL у реляційних базах даних. Найбільшого поширення набули три подібні між собою технології WebSQL, W3QS і W3QL.

1. Web Mining Research: A Survey, www.cs.kuleuven.ac.be/~dtai/publications/files/33042.ps.gz. 2. Ашманов И. Информация и знания: невидимая грань., http://newasp.omskreg.ru/intellect/f5.htm. 3. Alberto O. Mendelzon, George A. Mihaila, and Tova Milo. Querying the World Wide Web. Int. J. on Digital Libraries, l(l):54–67, 1997. 4. Kautz H., Selman B. and Shah M. The hidden web. AI magazine, 18(2): 27–36, 1997. 5. Query languages for the WWW, www.db.fmi.uni-passau.de/uni/WS97-98/Seminar/Ausarbeitungen/optimization.ps. 6. Informational Retrieval on the Web, www.trl.ibm.co.jp /kobayashi00information.pdf. 7. Mining the Link Structure of the Web www.almaden.ibm.coman /chakrabarti99mining.pdf. 8. Mladenic D. Text-learning and related intelligent agents // IEEE Intelligent Systems. 14(4):44–54, 1999. 9. Baeza-Yates R. and e. Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Longman Publishing Company, 1999. 10. Borges J. and Levene M. Data mining of user

navigation patterns // In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA, pages 31–36, 1999. 11. David Konopnicki. The w3ql query language and the w3qs system. Master's thesis, The Technion – Israel Institute of Technology, 1996. 12. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web // In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997. 13. Journey to the Internet's Unknown Regions, http://www.newsfactor.com/perl/story/17418.html. 14. http://www.sims. berkeley. edu/research/projects/how-much-info/internet.html.

УДК 681

О.Я. Тарас
Національний університет "Львівська політехніка",
кафедра "Інформаційні системи та мережі"

ОГЛЯД ІСНУЮЧИХ АЛГОРИТМІВ ТА МОДЕЛЕЙ ПОШУКУ У WEB

The review of algoritmes and searching modes are represented in this artickle.

Подано огляд алгоритмів та моделей пошуку, що застосовуються в пошукових системах.

ВСТУП

Пошукові і не тільки пошукові системи Інтернет настільки популярні сьогодні, що люди проводять години, обговорюючи переваги і недоліки, алгоритми та програми для пошуку повнотекстової інформації на тих або інших носіях. І весь цей час не вщухають гарячі суперечки між фахівцями та користувачами. Перші – прихильники суто "механічних" машин, пошукових систем, що обчислюють строгі логічні запити і підтримують усікання слова праворуч "зірочкою"; вони переконані, що краще всяких алгоритмів сформулюють, що ж їм потрібно знайти. Інші – навпаки, намагаються віддати алгоритмам розвідувача всі магічні перетворення вихідного запиту і не замислюватися про те, що ж там відбувається усередині. Обидві точки зору мають право на існування. Ми ж розглядатимемо цю проблему з позиції перших.

Бурхливий ріст обсягу інформації в Інтернет робить пошук незамінним методом доступу до цієї інформації. Можна виділити дві основні форми пошуку в Інтернет:

- *Використання пошукових систем*, що збирають відомості про ресурси, доступні в Інтернет, і організовують пошук за цією інформацією, як за повнотекстовою базою даних. Прикладами таких систем є – Altavista, Google, Яndex тощо.
- *Використання Інтернет-каталогів*, у яких інформація про обрані ресурси Інтернет класифікована за тематичними ознаками. Такі каталоги існують не тільки в електронному виді (List.Ru або Yahoo!), але також видаються і у вигляді друкованих видань – таких як, наприклад, "Жовті сторінки Інтернет''.