УДОСКОНАЛЕННЯ ПРОЦЕСІВ ЖИТТЄВОГО ЦИКЛУ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

УДК 002.5.004

М.С. Бурматова, М.В. Оленін

АНАЛІЗ ВИМОГ ДО АВТОМАТИЗОВАНИХ МЕТОДІВ ВИЛУЧЕННЯ ДАНИХ ПРО ОДНОТИПНІ ОБ'ЄКТИ З WEB-ПРОСТОРУ

Національний авіаційний університет
OleninMV@livenau.ua, burmatova.mariia@ukr.net

Проаналізовано 5 методів вилучення інформації про однотипні об'єкти з простору Web на відповідність поставленим вимогам: метод Sunny, методи обгорток, методи автоматизованого вилучення, синтаксичні методи та метод обробки Web-сторінки як текстового документу. Доведено найбільшу відповідність вимогам методу Sunny та методу обробки Web-сторінки як текстового документу.

Проанализированы 5 методов извлечения информации про однотипные объекты из пространства Web на соответствие поставленным требованиям: метод Sunny, методы оберток, методы автоматизированного извлечения, синтаксические методы и методы обработки Web-страницы как текстового документа. Доказано наибольшее соответствие поставленным требованиям метода Sunny и методов обработки Web-страницы как текстового документа.

The5 following listed Web-data mining methods are analyzed for the correspondence to the set requirements: method Sunny, wrappers method, automatic extraction method, syntaxic method, text mining method. Sunny and text mining methods are proved to be the most corresponding to the set requirements.

Ключові слова: вилучення даних, метод обгорток, автоматизоване вилучення, синтаксичні методи вилучення

Вступ

Стійкий прогрес розвитку комп'ютерних інформаційних технологій минулих трьох десятиліть призвів до появи великої кількості потужних комп'ютерів, програмного та апаратного забезпечення для зберігання та обробки даних. Ці технології зробили доступними користувачам величезну кількість баз даних та інших сховищ інформації для пошуку та вилучення з них інформації, а також для аналізу даних.

Завдяки цьому розвитку новітніх технологій WEB простір став найбільшим з відкритих джерел інформації сучасності. За статистикою 2009 року 93% нової інформації світу зберігається в електронному вигляді і є в тій чи іншій мірі доступною користувачеві. Окрім цього інформація у Web охоплює майже всі можливі теми і існує майже у всіх доступних формах (таблиці, текст, графічна інформація, відео, звук). Вона є динамічною і постійно змінюється.

Зі зростанням об'ємів інформації у Web зростає і необхідність розвитку та вдосконалення засобів вилучення і обробки інформації. Web-простір за своїм характером він є мало організованою розподіленою системою і не має чіткої організованої структури, тому і централізованої системи для пошуку і обробки інформації у Web також не існує.

На даний час для пошуку і вилучення інформації з Web використовуються автоматизовані пошукові системи.

Пошукові системи розрізняються методами обробки пошукового запиту, методами обробки джерел інформації, галузями застосування.

Найпоширенішим типом пошукових систем є універсальні повнотекстові пошукові системи, що охоплюють дуже велику кількість джерел, але в той же час надають мінімальні можливості користувачу для звуження занадто «загальних» результатів пошуку до необхідного.

Якщо користувач шукає загальну інформацію про якесь конкретне явище чи об'єкт – це не проблема, але якщо користувач шукає об'єкт з певними заздалегідь відомими характеристиками результати пошуку зазвичай є незадовільними. В першу чергу це стосується однотипних об'єктів.

Однотипні об'єкти – об'єкти, що мають визначений набір атрибутів і обмежений певними величинами набір значень цих атрибутів. Ці об'єкти можна віднести до певних класів, тобто вони підлягають класифікації. До однотипних об'єктів для прикладу можна віднести такі класи об'єктів, як автомобілі,

літаки, об'єкти нерухомості, комп'ютери, mp3-плеєри і т.д. Чим ці об'єкти цікаві?

По-перше задача пошуку об'єкту за характеристиками, а не за назвою об'єкту є доволі поширеною задачею у Web: протягом останніх десяти років надзвичайного зріс відсоток електронної комерції, дуже великого поширення набули Internet-магазини, агенції нерухомості і виробники публікують свої товари на власних Web-сайтах.

По-друге природа цих об'єктів (наявність чітких атрибутів і характеристик, класифікованість) є цікавою з наукової точки зору.

Для пошуку однотипних об'єктів у Web застосовуються спеціалізовані пошукові системи, що дозволяють здійснювати пошук інформації не тільки за повнотекстовим запитом, але і за певними значеннями атрибутів цих об'єктів. Спеціалізовані системи не охоплюють всього спектру знань \ інформації і постійно вимагають розширення.

Переважна більшість спеціалізованих пошукових систем охоплюють доволі невелику частину Web-простору і вимагають ручного заповнення своєї бази даних або імпортування даних з набору заздалегідь визначених Web-сайтів , окрім цього багато з Web-сайтів після знаходження відповідного об'єкту просто перенаправляють користувача до одного з джерел інформації, яке може бути неповним.

Таким чином задача побудови дружньої користувачу автоматизованої системи вилучення інформації про однотипні об'єкти з Web-простору є актуальною.

Постановка завдання

Проаналізувавши недоліки і переваги існуючих засобів вилучення інформації про однотипні об'єкти з простору Web [1], до засобу вилучення інформації необхідно поставити наступні вимоги:

1. Заповнення Бази даних системи (безумовно окрім етапу навчання \ налаштування системи) має бути автоматизованим.

2. Кількість джерел інформації не має обмежуватися різноманітними списками довіри і система має обробляти максимально доступну кількість джерел інформації.

3. Система має дозволяти здійснювати пошук інформації по значенням атрибутів однотипних об'єктів.

4. Система в результаті обробки даних має надавати добре структурований результат.

5. Система повинна містити засоби очистки результатів від інформаційного спаму та неякісних даних, або мати можливість вбудовування подібних засобів.

6. Система має надавати семантично відповідний запиту результат, для цього система має допускати звуження \ спеціалізацію для різних типів об'єктів.

Далі розглянемо існуючі автоматизовані методи вилучення інформації про однотипні об'єкти з Web-простору на відповідність поставленим вимогам.

1. Експериментальне визначення особливостей даних у Web

Розглянемо докладніше основний на даний момент формат представлення документів HTML\xHTML (hyper text markup language – гіпертекстова мова розмітки) – формат для представлення документів, який використовує певні структурні елементи – теги.

У синтаксисі мова розмітки має декілька типів тегів, основні з них - теги форматування, теги структури та теги гіпертекстових посилань. Звернемо увагу на те, що HTML – це дуже гнучка мова і за її допомогою можна представити один і той самий вміст багатьма способами.

Окрім того не зважаючи на те, що HTML є похідною мовою від жорстко структурованого SGML (Standard Generalized Markup Language – Стандартна узагальнена мова розмітки), HTML дозволяє вживати при описі документу неповністю ієрархічні структури, наприклад списки li, lo і перетинати ієрархічні елементи, наприклад запис abc defefg є дозволеним в HTML.

HTML документ легко може бути перетворений на звичайний текстовий документ, але тоді він буде позбавлений своєї структури. Таким чином Web-документи мають наступні суттєві властивості, які варто врахувати при побудові системи вилучення інформації з них: наявність певної первинної структурованості, забезпеченої синтаксисом HTML, гнучкість методів представлення інформації, неповна ієрархічність елементів документу та нежорсткість синтаксису HTML.

Було проведене базове дослідження для визначення того, наскільки часто ті чи інші HTML теги вживаються "за призначенням" (теги форматування – для форматування зовнішнього вигляду, теги структури – для побудови логічної структури).

Для дослідження було проаналізовано набір типових сторінок із 100 Web-сайтів ріелторів України. Кожен з дослідників отримав свою порцію типових сторінок з Web-сайтів і певний обмежений набір базових тегів для дослідження (DIV, P, FONT, B).

За допомогою простих інструментів дослідник визначав яке функціональне навантаження несе кожен екземпляр досліджуваних тегів у документі. Тег міг нести наступне навантаження:

- Використовуватися як тег структури
- Використовуватися як тег форматування
- Нечітко визначене (змішане) використання (в дослідженні такі екземпляри відносилися до використання по призначенню)

Безумовно, перехід від використання тегу як тегу форматування до використання тегу як тегу структуризації є дуже умовним.

Тому для підвищення об'єктивності дослідження кожен дослідник мав створити умовну логічну схему документу відповідно до вмісту документу і дослідити структуру документу на відповідність логічній схемі. Наявность або відсутність відповідності між логічною і структурною схемою занотовувалася дослідником.

Результати дослідження представлені на рис 1.1.

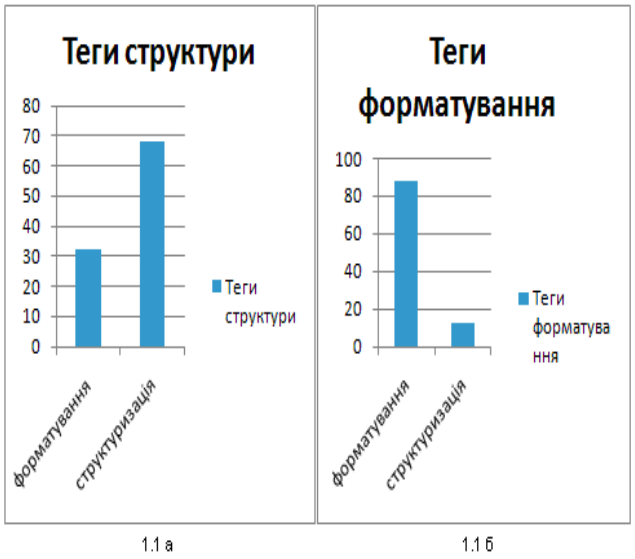
З рис.1.1 видно, що дуже великий відсоток (32) тегів структурування вживаються в якості тегів форматування.

Це нівелює їх роль елементів побудови логічної структури документу і не дозволяє повноцінно використовувати їх при розборі Web-документу. Окрім того значний відсоток тегів форматування (15) вживаються для структуризації документів.

В той же час ми бачимо, що дуже значний відсоток тегів вживається для структурування документу.

Таким чином на практиці виявляється, що Web-документ – це набір частково структурованих даних, що містить вміст та граматично визначені елементи структури та елементи форматування, які можуть бути нерозрізнюваними між собою і відповідно при вдосконаленні схеми документу можуть бути заміненими на інші теги без звернення уваги на тип тегу.

Очевидно, що теги, відповідальні за побудову структури Web-документу, мають безпосереднє відношення до вмісту документу і визначаються семантикою вмісту.



1.1 а 1.1 б

Рис.1.1: а - діаграма використання тегів структури для структурування документу (1), для форматування (2); б - діаграма використання тегів форматування для структурування документу (1), для форматування (2)

Наявність певної структури, прив'язаної до семантики документу, є головною перевагою Web-документів над звичайними текстовими документами. Неповнота і недосконаліть цієї структури вимагає перетворення документу і подальшої його структуризації.

Додатково було проведено експериментальне дослідження в галузі об'єктів нерухомості на прикладі набору з 3920 Web-сайтів основних ріелторів України. Webсайти були проаналізовані на предмет наявності повторюваної структури, достатньої кількості інформації для автоматичного вилучення, наявності інформаційного спаму.

Дослідження було проведене на двох етапах: на першому людина-оператор аналізувала Web-сайт шляхом перегляду його зовнішнього вигляду, HTML-коду і в разі очевидної неструктурованості, замалої кількості інформації, завеликої кількості спаму відсіювала Web-сайти.

На другому етапі для набору Web-сайтів, що залишилися вручну створювалися шаблони вилучення, що могли бути створені лише у випадку якісного представлення інформації. Результати представлені у таблиці 1.1.

Таблиця

Результати аналізу Web-сторінок з основних ріелторів України на придатність до вилучення інформації

Кількість Web-сайтів	Кількість у відсотках, %	Опис
1304	33,27	Придатні
1308	33,37	Непридатні у зв'язку з очевидною відсутністю спільної структури
201	5,13	Непридатність у зв'язку з неструктурованістю, визначена після повторного більш глибокго аналізу
452	11,53	Замала кількість інформаціїі
547	13,95	Спам
108	2,76	Непридатні за інших причині
3920	100	Загалом

Як видно з таблиці більш ніж 30 відсотків інформації були відсіяні за повної невідповідності вимозі структурованості, при подальшому більш глибокому аналізі вмісту з тих, що залишилися після першої фази аналізу були відсіяні ще 5 відсотків.

Через замалу кількість інформації були втрачені ще 11 відсотків інформації.

Таким чином втрати джерел корисної інформації складають неприпустимі 67% від загального об'єму джерел. Велику частину цієї інформації насправді можна видобути, застосовуючи більш гнучкі методи вилучення.

Для з'ясування взаємозв'язку між послідовністю опису атрибутів документу було проведено третій експеримент:

Користувач на наборі з 300 Web-сторінок, кожна з яких містила не менше 4-х об'єктів нерухомості, за допомогою спеціально розробленої допоміжної програми виділяв окремі об'єкти на сторінці і в середині них значимі фрагменти об'єктів, такі як:

- ціна
- адреса
- площа
- кількість кімнат
- тип нерухомості
- тип операції
- поверх будинку
- контактна інформація

Вимоги розмітки були наступними:

- користувач мав виділити всі наявні значимі фрагменти як мінімум для трьох об'єктів
- користувач не мав права виділяти лише частину з наявних фрагментів для окремих об'єктів

Після розмітки кожної сторінки інформація про послідовність значимих фрагментів виділялася у таблицю бази даних. Після заповнення бази даних було зроблено їх статистичну обробку для з'ясування наявності взаємозв'язків між фрагментами документу.

Експеримент виявив, що послідовність значимих частин опису об'єкту, які відповідають атрибутам об'єктів, не є фіксованою, жорстких послідовностей немає.

Таким чином збереження всього дерева документу при з'ясуванні чи описує наданий фрагмент документу якийсь з атрибутів окремого об'єкту не є доцільним, збільшує об'єм розрахунків і може навіть погіршити результат розрахунків, якщо автоматизована система вилучення знайде і буде використовувати несуттєву закономірність.

2. Методи обробки даних з Web

Розглянемо які методи можна застосувати для вилучення інформації з Web. Серед них варто виділити наступні:

- Перетворення Web-документу у текстовий документ і використання синтаксичних, морфологічних, семантичних методів розбору тексту Text Mining (вилучення даних і знань з природномовних текстів) для вилучення необхідної інформації
- Використання синтаксичних методів для вилучення інформації з Web-сторінки без її перетворення у текстовий документ (власне Web-сторінка розглядається як текстовий документ)
- Використання семантичних методів для вилучення інформації з Web-сторінки без її перетворення у текстовий документ (власне Web-сторінка розглядається як текстовий документ)
- Перетворити на набір взаємозв'язаних елементів і застосувати методи data mining для вилучення необхідної інформації
- Змішані методи

Розглянемо докладніше доцільність використання кожного з вищезгаданих методів для вилучення інформації про однотипні об'єкти.

2.1. Обробка Web-сторінки як текстового документу

Синтаксичне вилучення інформації – розділ математичної та обчислювальної лінгвістики, в якому визначається набір правил для обробки, перетворення та ранжування елементів тексту – словоформ (слів та словосполучень), які виконуються за допомогою лінгвістичних процесорів (Natural Language Processors). Центральної функцією мовних процесорів є граматичний розбір (parsing).

Програми граматичного розбору (parsers) застосовують в якості довідкових даних формальні граматики і словники тієї мови, тексти до-якої слугують об'єктом аналізу або синтезу.

Як формальні граматики використовуються розширені граматики безпосередніх складових (контекстно-вільні граматики), трансформаційні граматики, граматики розширених мереж переходів, що є системами граматик безпосередніх складових, і ін.

Як формальні словники використовуються прикладні (інженерні) варіанти толково-комбінаторних словників, т. тобто спеціальні форми семантико-синтаксичних словників, що мають докладну інформацію про варіантні форми слів, про їх семантику і про поєднувальну можливості на лексичному, семантичному і синтаксичному рівнях з урахуванням морфологічних обмежень. Звичайні мовні процесори містять морфологічну, синтаксичну, семантичну (або синтактико-семантичну) і словникову компоненти (підсистеми програм і даних), кожна з яких реалізує динамічну модель мови на відповідному рівні.

Знання часто представляються у вигляді т. зв. фреймів - мовних моделей певних фрагментів дійсності або семантичних мереж і утворюють т. н. бази знань, що зберігаються в комп'ютері.

Ці функції використовуються також і як засобу розкриття неоднозначностей, поновлення Еліпсів, встановлення анафоричних зв'язків у тексті і в інших складних випадках лінгвістичного аналізу.

Серед головних переваг методів Text Mining – високий потенціал – при повній реалізації система теоретично може замінити людину у таких галузях як переклад, редагування текстів, вилучення інформації з текстів. Таким чином повна реалізація дозволить здійснювати пошук інформації по значенням атрибутів однотипних об’єктів та здійснювати надання інформації користувачеві у найбільш зручній формі – у вигляді природно мовного тексту, який добре відповідає користувацькому запиту.

Окрім того кількість джерел інформації для такої системи є найбільшою серед усіх розглянутих систем. Серед недоліків головними є надвелика трудоємність реалізації, що окрім підготовки самих методів вилучення вимагає підготовку словників, фреймів, семантичних мереж, а також прив’язаність до лексики і граматики конкретної мови, що означає що вимога автоматизованого заповнення бази знань самої системи не може бути дотриманою.

Окрім того при застосуванні методів до задачі вилучення інформації про однотипні об’єкти з простору Web при перетворенні Web-документів у текстовий документ вилучається лише вміст тегів. При цьому втрачаються елементи форматування (кольори, картинки, спеціальні відступи і т.п.).

Відповідно структура документу зазнає певних змін, які можуть бути невидимими для кінцевого користувача, але як визначило попередньо проведене дослідження – переважна частина з яких є прив’язаною до семантики тексту.

Таким чином головна перевага Web-документу над звичайним текстовим документом втрачається.

Очистка даних від інформаційного спаму, отриманих системою такого типу, вимагає розробки додаткових методів семантичного аналізу тексту і є додатковою трудомісткою задачею.

2.2. Синтаксичні методи перетворення Web-документу

Синтаксичні методи перетворення Web-документу – це методи, що частково застосовують технології знаходження послідовностей у неструктурованих даних іноді в комбінації з методами Text Mining, розглядаючи Web-документ як неструктуровану послідовність і залежно від конкретного методу символів\слів\уривків тексту\тегів. Ці методи – це переважно засновані на застосуванні шаблонів, попередньо створених користувачем на основі «ручного» аналізу тексту. Іноді застосовуються методи кластеризації та ранжування і індексації уривків тексту в комбінації з механічно заповненими словниками, шалонами і фреймами послідовностей. Хоч ці методи і не виключають з розгляду HTML структуру документу, вони розглядають її не як засіб організації документу, а просто як

додаткову послідовність текстової інформації. Для вдалого функціонування ці методи вимагають великої кількості попередньої обробки джерел інформації, створення шаблонів, а деякі (більш універсальні) вимагають створення словників і фреймів. В будь-якому разі про автоматичне заповнення бази знань системи не йдеться.

В той же час системи такого типу дозволяють пошук інформації по значенням атрибутів однотипних об'єктів, результат на запит користувача є добре структурованим і відповідний семантиці запиту, окрім цього дані зазвичай не містять інформаційного спаму. Методи підходять для локальних реалізацій при обробці даних, послідовність (структура) яких лишається незмінною протягом великої кількості часу, що є дуже серйозним обмеженням джерел інформації.

2.3. Data Mining методи перетворення Web-документу

Вилучення даних з Web (Web content mining) - це автоматизований процес знаходження, вилучення та інтеграції корисних даних отриманих з Web-документу, що включає зміну структури та вмісту документу до форми придатної до обробки комп'ютером.

Серед методів вилучення інформації є методи вилучення структури документу, вмісту документу, використання документу. З вищезазначених нас цікавлять автоматизовані методи вилучення вмісту документу.

Існує два типи Data Mining методів вилучення інформації про однотипні об'єкти з простору Web (методів вилучення вмісту документів):

Метод використання обгорток (Wrappers) – розбір структури документу шляхом помічення частин документу та створення карти документу

Автоматизоване вилучення даних – розбір структури документу шляхом надання набору позитивних прикладів або шляхом надання однієї сторінки з набором позитивних прикладів

Розглянемо докладніше кожен з методів.

2.3.1. Метод використання обгорток

В методах обгорток використовується машинне навчання для створення правил вилучення інформації[4]. Це відбувається наступним чином:

1. Користувач помічає цільові поля (частини коду, що містять інформацію яку потрібно вилучити) в наборі навчальних сторінок

2. Система знаходить набір правил, що описують кожне з цільових полів

3. Правила застосовуються для вилучення інформації з інших сторінок з тим самим форматом, що й ті, на яких проводилося навчання.

Розрізняють наступні види методів використання обгорток:

1. Методи ієрархічного навчання, в яких кожне цільове поле (кожен атрибут однотипного об'єкту) вилучається незалежно від інших і кожен є ізольованим на своєму рівні ієрархії Web-документу.

2. Методи спільного навчання, в яких ієрархія документу не обмежує процес вилучення кожного атрибуту об'єкту і атрибути вилучаються разом для об'єкту в цілому.

Зазвичай набір вилучених правил у методах ієрархічного навчання складається з правила початку і правила кінця атрибуту, які однозначно описують положення кожного атрибуту в коді.

Для методів спільного навчання набір вилучення правил окрім правил початку і кінця атрибуту містить правила початку і кінця кожного об'єкту і правила, що стосуються атрибутів, є відносними до правил меж об'єкту (вони однозначно описують положення атрибуту в частині коду, що визначає окремий об'єкт).

Навчання в методах використання обгорток відбувається поетапно: на кожній ітерації система створює набір ідеальних правил що якомога більше позитивних ситуацій і жодної негативної, як тільки позитивний приклад стає описаним правилом він видаляється з навчальної вибірки. Навчання відбувається до тих пір поки всі позитивні ситуації не охоплюються правилами.

Серед переваг систем обгорток можна виділити наступні: системи обгорток мають відносно нескладний алгоритм пошуку правил; системи надають доволі високу якість результату на добре підготовленому наборі даних; є можливість пошуку по значенням атрибутів об'єктів; результат пошуку є добре структурованим і семантично відповідає пошуковому запиту користувача; методи очистки від інформаційного спаму не є потрібними у зв'язку з обмеженою кількістю джерел інформації; заповнення бази знань системи є автоматизованим процесом, що вимагає нескладного навчання користувачем.

Недоліки систем обгорток є наступними: створення міток здійснюється для кожного окремого Web-сайту, з якого необхідно вилучати

інформацію; дуже складна підтримка в разі частих змін структури Web-сайту; кількість джерел інформації обмежена лише тими, в яких є чітка повторювана структура для представлення однотипних об'єктів. Будь-яка зміна в Web-сайті може зробити існуючий набір правил невірним і для виправлення необхідне нове навчання і нове ручне надання міток користувачами

Системи обгорток непридатні для вилучення інформації з неструктурованих сайтів, або сайтів з замалою кількістю інформації і не зважаючи на численні переваги системи обгорток потребують наявності достатньої кількості добре структурованих сторінок в якості вхідної інформації, що видається неможливим за теперішньої ситуації на українському Web-просторі.

2.3.2. Автоматизоване вилучення

В методах автоматизованого вилучення в інформації з набору Web-сторінок отримання знань відбувається майже без залучення користувача. Алгоритм, отримавши набір позитивних прикладів, покроково створює набір правил для кожної з цих сторінок[3]. Це відбувається наступним чином:

1. Система отримує набір сторінок з одним або більше позитивними випадками або сторінку з декількома позитивними випадками.
2. Система вважає першу цілу вхідну сторінку обгорткою.
3. Далі система починає перевизначати обгортку, шляхом знаходження і розв'язку невідповідностей між обгорткою і кожною сторінкою.

Невідповідністю вважається ситуація в якій елементи сторінки не відповідає граматиці обгортки. Існують наступні типи невідповідностей:

Невідповідність вмісту тегів (вказує на наявність полів з даними)

Невідповідність тегів (вказує на опціональні елементи, ітератори)

Ітератори на відміну від опційних елементів з'являються на початку або в кінці повторюваного шаблону і визначаються наступним чином:

Система знаходить відносну позицію невідповідності і визначає деяких кандидатів у повторювані шаблони, шукаючи ці шаблони далі.

Система порівнює кожен з кандидатів з попередньо знайденими шаблонами для підтвердження правильності припущення.

При підтвердженні припущення для кожного з наступних шаблонів обгортка розширюється, при непідтвердженні система бере інший шаблон-кандидат у ітератори.

Методи автоматизованого вилучення мають наступні проблему: Об'єм розрахунків алгоритму знаходження відповідності експоненціально пропорційний довжині вхідного рядку (фактично розміру документу, що аналізується). Для зменшення об'єму розрахунків використовуються наступні евристичні стратегії: зменшення об'єму дослідження, зменшення бектрекінгу

Шаблони не можуть бути розмежовані ні на початку ні на кінці опціональними елементами (таким чином зменшується складність правил)

В порівнянні з методом обгорток системи автоматизованого вилучення мають перевагу відсутності ручної розмітки вхідних сторінок (в той же час вимога надання набору позитивних сторінок з однаковою розміткою для автоматичного вилучення залишається, що означає, що кількість джерел інформації є обмеженою). Заповнення бази знань системи відбувається автоматично

Недоліки систем автоматизованого вилучення перераховані далі: обгортка будується не для даних з Web-сторінки, а для всієї Web-сторінки, що призводить до великої кількості проблем пов'язаних з необхідністю відсіювати неважливу інформацію та спам, які присутні на переважній більшості сторінок. Системи автоматизованого вилучення мають проблеми з визначенням опціональних елементів та диз'юнктивного опису даних. Відсутність можливості задавати імена атрибутам у вилучених даних призводить до необхідності створювати засоби інтеграції даних вилучених з численних Web-сайтів, неможливості пошуку інформації по атрибутам об'єктів, неможливості отримання структурованого та семантично відповідного запитам результату.

Системи автоматизованого вилучення непридатні для вилучення інформації з неструктурованих сайтів, або сайтів з замалою кількістю інформації.

2.4. Метод послідовного перетворення Web-сторінки з наступним вилученням інформації з неї

Розлянуті вище автоматизовані методи вилучення інформації розраховані на те, що Web-документ має чітку, повторювану несуперечливу структуру і в ньому достатньо

інформації для навчання. На жаль переважна більшість джерел інформації про однотипні об'єкти з українського простору Web не відповідають цим вимогам як це доведено у експерименті з розділу 1.

Метод Sunny може використовувати неструктуровані або слабко структуровані документи для навчання та вилучення інформації про однотипні об'єкти. Він заснований на наступних особливостях Web-документів, що містять опис однотипних об'єктів:

- Наявність структури, зумовленої синтаксисом HTML
- Невідповідність тегів тим функціям, які вони використовують (теги форматування можуть використовуватися в якості тегів структури і навпаки)
- Наявність певних синтаксичних констант (сталих виразів, слів, конструкцій) притаманних лише певним галузям однотипних об'єктів
- Несильні семантичні зв'язки у дереві документу між вузлами дерева, які описують окремі атрибути об'єктів

Метод Sunny є методом вилучення інформації з контрольованим навчанням. Під час навчання від користувача системи вимагаються наступне: за допомогою спеціально розроблених для цього засобів завантажити документ\набір документів з локального, мережевого сховища або з Web-простору і розмітити кожен з документів, виділяючи наступне:

- Межі окремих однотипних об'єктів
- Значення атрибутів для всіх окремих однотипних об'єктів
- Окрім цього користувач має задати константи предметної області та специфічні для предметної області записи.

Система в свою чергу здійснює наступне:

- Під час завантаження документу здійснює структурне перетворення документу
- Після виділення меж окремих об'єктів здійснює навчання для визначення релевантної та нерелевантної інформації
- Після розмітки атрибутів для кожного окремого об'єкту здійснюється структурне перетворення та навчання системи.

Для навчання інформація про кожен розпізнаний за допомогою користувача атрибут об'єкту, вміст якого описується окремим тегом, заноситься у таблицю бази даних.

Для навчання використовується метод побудови дерев рішень C4.5.

Власне автоматизоване вилучення інформації про однотипні об'єкти відбувається на наступних кроках:

1. Для зменшення кількості розрахунків, покращення їх якості з документу видаляються фрагменти, які не є описом однотипних об'єктів. Для цього використовуються методи ранжування вмісту Web-сторінок. На жаль ранжування сторінок не є предметом даної статті дослідження і детально розглянуте не буде. Зауважимо лише, що для ранжування використовуються такі фактори як наявність ключових слів (позитивних і негативних) в тексті, їх положення та щільність.
2. HTML-документ позбавляється суперечливих логічно незакінчених елементів і приводиться до чіткої ієрархічної структури, тобто перетворюється в формат XHTML. При цьому теги розмітки документу перетворюються на функціональні теги структури документу. Таке перетворення називається структурним перетворенням, а отримана модель – структурною моделлю.
3. До структуризація структурної моделі шляхом виділення певних скорочень, загальноприйняті термінів ті їх послідовностей (синтаксичних констант), виділення певних стандартних (для аналізованої галузі) методи представлення інформації, виділення розділювачів, які роз'єднують не зв'язані між собою фрагменти тексту. Це перетворення називається синтаксичним перетворенням.
4. Знаходяться повторювані послідовності у структурі і вмісті документу. Знайдені послідовності\шаблони є фактично знаннями, які можна занести до сховища даних. Для знаходження послідовностей використовується алгоритм вилучення записів даних з Web MDR [2].
5. Вилучється інформація з XHTML дерева у БД зі збереженням інформації про найближчі взаємозв'язки у древі
6. Використовується метод дерев рішень C4.5 для подальшої обробки інформації в базі даних і вилучення закономірностей опису даних.

Метод Sunny є автоматизованим методом вилучення інформації з Web-простору, що не обмежує джерела інформації лише добре структурованими документами, він дозволяє вилучити інформацію про окремі атрибути об'єктів. Окрім того цей метод дозволяє надати структурований результат завдяки використанню релевантної бази даних. Метод дозволяє використовувати засоби очистки від

нерелевантної інформації перед та після вилучення інформації. Метод надає можливість «спеціалізації» на кожному окремому класі однотипних об'єктів.

Висновки

Було проаналізовано 4 основні групи автоматизованих методів вилучення інформації про однотипні об'єкти та метод системи Sunny, що не належить до жодної з перерахованих груп, на предмет їх відповідності поставленим у [1] вимогам.

Найбільшу кількість джерел інформації може обробляти метод обробки Web-сторінок як текстових документів, другим є метод Sunny, найменш універсальними є методи автоматизованого вилучення, методи обгорток та синтаксичні методи.

Метод Sunny, методи обгорток та обробки Web-сторінок як текстових документів, синтаксичні методи дозволяють здійснювати вилучення інформації для кожного окремого атрибуту однотипного об'єкту і відповідно дозволяють надавати добре структурований результат пошуку на користувацький запит, що семантично відповідає запиту.Кожна з систем вимагає окремого засобу відсіювання інформаційного спаму.Метод обробки Web-сторінок як текстових документів є надзвичайно трудомістким і не є реалізованим в повній мірі для мов присутніх в українському Web- просторі.

Шляхом аналізу існуючих методів вилучення інформації про однотипні об'єкти з Web було доведено, що більшість методів (окрім методу Sunny та частково методу обгорток) в повній мірі не використовують особливостей однотипних об'єктів, втрачають багато інформації та галузь їх застосування є доволі вузькою. Таким чином найбільш відповідають поставленим вимогам метод Sunny та метод обробки Web-сторінок як текстових документів, який на даний момент реалізувати в повній мірі майже неможливо.

Список літератури

1. *Бурматова М.С., Оленін М.В.* Аналіз сучасних пошукових систем на предмет їх придатності для пошуку і вилучення інформації про однотипні об'єкти з Web-простору: матеріали міжнародної науково-технічної конференції УкрПрог 2010, Київ, Україна – К., 2010

2. Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn. 2004 Nov;4(6):795-803. [Electronic resource] : proceedings. – Mode of access: WWW.URL: http://www.multifactordimensionalityreduction.org/– Last access: 2010.

3. *Liu, B., Grossman, R., Zhai, Y.* Mining Data Records in Web Pages. KDD-03, 2003.

4. Zhao, H., Meng, W., Wu, Z., Raghavan, V.,Yu, C. Fully automatic wrapper generation for search engines. WWW-05, 2005

Відомості про авторів:

Бурматова Марія Сергіївна, менеджер проектів, Infopulse LLC, наукові інтереси – data mining, data storage solutions
E-mail: burmatova.mariia@ukr.net

Оленін Михайло Вікторович, Національний авіаційний університет, кафедра інженерії програмного забезпечення, доцент, кандидат технічних наук; наукові напрями – інтелектуальні системи
E-mail: OleninMV@livenau.ua

Стаття надійшла до редакції 22.06.2010