DOI https://doi.org/10.15407/usim.2018.04.0032
УДК004.9

В.В. ЗОСІМОВ канд. техн. наук, доцент,
Миколаївський нац. ун-т ім. В.А. Сухомлинського,
м. Миколаїв, 54030, вул. Нікольська, 24, тел.: (0512) 37-88-09
zosimovvv@gmail.com

КОМПЛЕКСНИЙ ПІДХІД ДО ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ОБРОБКИ ВЕБ-ДАНИХ НА ОСНОВІ СЕМАНТИЧНОЇ РОЗМІТКИ

Розроблено систему комплексного оперування даними в мережі Інтернет, яка надає користувачу ефективні, зручні та прості у використанні інструменти обробки веб-даних на всіх етапах взаємодії з семантичною павутиною від створення веб-ресурсів до пошуку інформації. В основу системи покладено нову предметно орієнтовану мову оперування веб-даними та словник семантичної розмітки корпоративних веб-ресурсів.

Ключові слова: *пошук інформації, модель ранжування, метапошукова система, семантична розмітка, онтології, структура веб-ресурсу, видобування даних, предметно орієнтована мова, система керування вмістом.*

Вступ

Поняття *пошук інформації в мережі Інтернет* в науковій літературі зводиться, як правило, до вивчення алгоритмів роботи пошукових систем. До них відносять збір та індексацію інформації з веб-ресурсів, пошук веб-сторінок, які відповідають ключовим словами пошукового запиту і ранжування отриманих результатів в порядку релевантності пошуковому запиту [1].Пошук та аналіз текстової інформації досліджували видатні вітчизняні та зарубіжні вчені: Ньюелл А., Люгер Д. Ф., Фостер Д. М., Анісімов О.В., Поспєлов Д. О., Попов Є. В., Широков В. О.

Однак мало хто розглядає процес пошуку інформації в Інтернеті з точки зору користувача, для якого отримання списку веб-ресурсів, релевантних його запиту, є лише першим кроком до отримання шуканої інформації. Існує ряд факторів, які мають значний вплив на швидкість і зручність пошуку інформації з точки зору користувача. Їх можна розділити на дві групи:

Якість пошукової видачі:

- інструменти взаємодії користувача з пошуковою видачею (фільтри, сортування, вибір методів і моделей ранжування);
- візуальне представлення конкретного веб-ресурсу в пошуковій видачі, тобто якість і інформативність його опису;
- ранжування результатів пошуку з урахуванням контексту введеного пошукового запиту. Тут маються на увазі методи, які дозволяють пошуковій системі зрозуміти, що шукає користувач — товар, послугу, інформаційні статті або наукові публікації.

Якість веб-сторінки, представленої в результатах пошуку:

- зручність навігації за веб-ресурсом —впливає на те, наскільки швидко користувач може розібратися зі структурою веб-ресурсу і відшукати на ньому необхідну інформацію або здійснити дію, наприклад, замовити послугу;
- зручність подання інформації на вебсторінці (шрифт, розмір тексту, кольору, розташування елементів, інше стильове оформлення);

▪ наявність реклами, яка ускладнює сприйняття інформації (наскільки вона нав'язлива, її кількість, розташування, можливість відключити);

▪ унікальність і якість інформації —чи є вона унікальною, або це вільний виклад (рерайт) існуючої на інших веб-ресурсах інформації.

Пошук інформації з точки зору користувача та пошукової системи

Пошук інформації в мережі Інтернет з позиції користувача та з позиції пошукової системи це два різних процеси, які мають різні цілі і результати.

Пошук інформації з огляду на пошукову систему, полягає в знаходженні серед безлічі представлених в мережі веб-сторінок таких, які відповідають ключовим словами, введеним в рядку запиту. Потім відбувається ранжування отриманих результатів на підставі закритих від користувача алгоритмів, на роботу яких він ніяк не може вплинути. Тобто ранжування результатів проводиться за однаковими моделям без урахування контексту пошуку, індивідуальних потреб користувача і предметної області. Додаткові опції, пропоновані при розширеному пошуку не можна вважати інструментом впливу на ранжування результатів, оскільки вони є фільтрами, що відсіюють частину результатів пошуку [2]. На цьому завдання пошукової системи вважається виконаним.

Для користувача метою пошуку інформації є отримання конкретної інформації з певної предметної області в певному контексті. На етапі перегляду результатів пошуку користувач може оцінити ймовірність присутності на веб-ресурсі необхідної інформації за стислим описом сторінки з мета-тега <description> і фрагментів тексту сторінки, що містять одне з ключових слів. Достовірність такої оцінки ускладнює той факт, що ці мета-теги заповнюють фахівці зі штучного просування веб-ресурсів, які знають як зробити їх привабливими в першу чергу для пошукових систем, а потім — для користувача. Інформація в мета-тегах не завжди достовірно відображає інформаційний зміст веб-сторінки, і складається для залучення на веб-ресурс потенційних клієнтів.

Після прийняття рішення про те, що розміщена на веб-ресурсі інформація може бути корисною, користувач переходить за посиланням. На цьому етапі перед користувачем постає нове завдання пошуку необхідної інформації в межах одного веб-ресурсу. Значний вплив на швидкість отримання результату і на можливість в принципі його отримати впливає структура, система навігації і візуальне оформлення веб-ресурсу.

Для успішного досягнення цілей пошуку, користувачеві необхідно отримати прості і зрозумілі в застосуванні інструменти, які дозволяють ефективно долати описані труднощі, що виникають при пошуку інформації.

Можливі шляхи підвищення ефективності пошуку інформації

Виходячи з вищесказаного актуальним завданням є розробка методів персоналізації результатів роботи пошукових систем шляхом надання користувачу інструментів управління пошуковою видачею, а також використання нових моделей ранжування[3], заснованих на суб'єктивних для кожного користувача параметрах.

На етапі роботи з пошуковою видачею необхідно надати користувачеві інструменти налаштування параметрів відображення списку веб-ресурсів, починаючи з вибору елементів вмісту, які відображатимуться, і закінчуючи візуальним оформленням кожного елемента списку. Потужним інструментом для підвищення ефективності пошуку інформації є використання пошукових агентів, тому в системі мають буті реалізовані програмні засоби розробки пошукових агентів, доступні широкому колу користувачів без досвіду в галузі розробки веб-додатків.

На етапі пошуку інформації на конкретному веб-ресурсі необхідно реалізувати новий підхід до відображення корисної для користувача інформації в тому вигляді, який дозволить йому максимально швидко її сприймати, витрачаючи при цьому мінімум часу на ознайомлення зі структурою, системою навігації та візуальним оформленням веб-ресурсу. Ця концепція пе-

редбачає уніфікацію відображення інформації для часто використовуваних типів веб-ресурсів (сайт компанії, інтернет-магазин, сайт новин, блог) на основі семантичних шаблонів.

Розробка таких методів передбачає широке використання семантичної розмітки в коді веб-сторінок, що дозволить застосовувати методи машинної обробки представленої на них інформації.

Розробка концепції семантичної павутини стала наступним кроком розвитку глобальної мережі. Розміщена в мережі Інтернет інформація зручна для розуміння людиною. Семантична павутина була розроблена для того, щоб зробити інформацію придатною для автоматичного аналізу та синтезу висновків. Незважаючи на явні переваги застосування даної технології, вона не набула значного поширення в веб-середовищі [4].

Інтеграція бізнес-процесів України в середовище закордонних партнерів диктує необхідність розвитку сфери електронної комерції як необхідної умови існування сучасних компаній.

Значні результати досягнуті в розробці моделей семантичної розмітки інтернет-магазинів як основного інструменту ведення електронної комерції. В той же час досить мало уваги приділяється електронному ринку послуг, а саме структурним та семантичним стандартам розробки корпоративних веб-ресурсів. В українському сегменті всесвітньої павутини лише незначний відсоток веб-ресурсів, розроблений з використанням стандартів семантичної розмітки. Така ситуація стала наслідком існуючих з самого зародження концепції семантичної павутини, проблем практичної реалізації та особливостей вітчизняного ринку розробки веб-ресурсів.

Для реалізації зазначеного підходу необхідно розв'язати наступні задачі:

- інтеграцію семантичної розмітки до існуючих веб-ресурсів;
- розробку нових веб-ресурсів із вбудованою семантичною розміткою;
- розробку методів видобування даних з веб-ресурсів з інтегрованою семантичною розміткою та без неї.

Система комплексного оперування даними в мережі Інтернет

В статті представлена системакомплексного оперування даними в мережі Інтернет (КОДІ) в межах концепції семантичної павутини, орієнтованої на вдосконалення методів пошуку інформації, створення веб-ресурсів з інтегрованою семантичною розміткою та програмних пошукових агентів.

Робота системи представлена на прикладі обробки даних, представлених на корпоративних веб-ресурсах українського сегменту всесвітньої павутини.

Далі подано загальну структуру системи та опис її окремих модулів.

Структурна схема системи КОДІ. В основу системи покладено нову предметно орієнтовану мову оперування веб-даними, що містить всі необхідні функції для видобування, збереження та відображення інформації, представленої на веб-ресурсах. Другим важливим компонентом системи є розроблений словник семантичної розмітки корпоративних веб-ресурсів, використання якого значно прискорює обробку даних.

Для забезпечення ефективної роботи технології необхідно дослідити вміст корпоративних веб-ресурсів, розробити загальну структуру та словник семантичних тегів для опису побудованої структури.

Загальна схема системи КОДІ представлена на рис. 1.

Для визначення структури корпоративних веб-ресурсів було проведено експеримент, в ході якого покроково було досліджено: загальну структуру, елементи навігації, інформаційні розділи.

В ході експерименту досліджено структуру 1000 корпоративних веб-ресурсів. Його аналіз та видобування даних здійснювалось за використання автоматичного парсера, реалізованого засобами розробленої в роботі мови оперування веб-даними.

Веб-ресурси для аналізу були автоматично відібрані з видачі пошукової системи Google на запит «Наша компанія». Таке формулювання пошукового запиту очевидно забезпечує високу

імовірність наявності в результатах пошуку саме корпоративних сайтів. Загалом було переглянуто 1284 результати пошуку, з них були відсіяні повторні посилання на вже оброблені веб-ресурси, дошки об'яв та сайти-агрегатори послуг.

Наступним кроком дослідження є побудова семантичного словника представлення корпоративних веб-ресурсів.

Дослідження існуючих рішень показало, що як базовий набір класів для опису семантичної структури корпоративних веб-ресурсів доцільно використовувати стандарти Good Relations, який є спеціалізованим розширенням найбільш поширеного словника Schema.org для електронної комерції.

Використання готових стандартів гарантує високий рівень сумісності розробленої технології з існуючими інструментами.

Цей стандарт містить необхідні класи для опису: контактні дані для компаній та співробітників, відгуки, продукцію, ціни, способи доставки та оплати.

Враховуючи базову онтологію, було додано нові класи згідно структури корпоративних веб-ресурсів.

Предметно-орієнтована мова оперування веб-даними. Будь-яка обробка даних, видобутих з веб-сторінок не є складним завданням і може бути ефективно реалізована існуючими засобами програмування. Нетривіальним є завдання саме видобування інформації з веб-ресурсів. Існує безліч підходів до створення веб-сторінок і використання кожного з них на виході генерує свій унікальний html-код. Тому сьогодні не існує ефективних програмних рішень, які дозволили б уніфікувати цей процес для всіх веб-ресурсів.

Сьогодні для отримання даних з веб-сторінки застосовуються спеціальні програми — грабери. Їх завдання — збір інформації з певних веб-сайтів за певними параметрами. Але для кожного конкретного веб-ресурсу необхідно писати окрему програму, яка зможе отримувати дані з урахуванням стильових, програмних і структурних особливостей кожного веб-ресурсу. Для вирішення завдання уніфікації процесу видобування розроблено предметно-

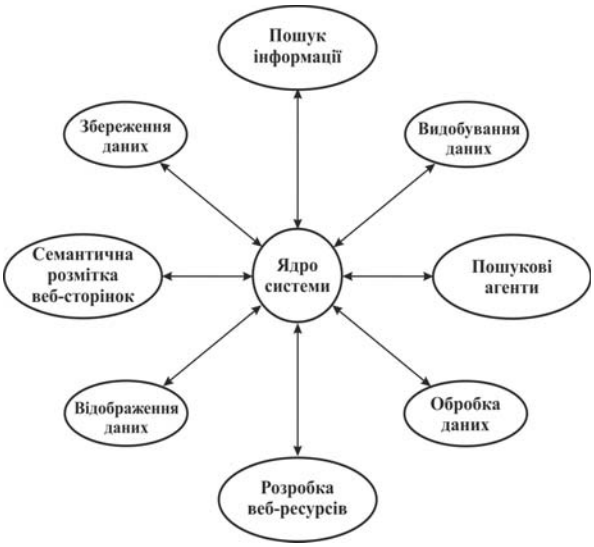

Рис. 1. Загальна схема системи КОДІ

орієнтовану мову (ПОМ) з широким функціоналом для оперування веб-даними [5-7].

Метою розробки такої мови було забезпечення системи зручними та інтуїтивно зрозумілими програмними засобами обробки веб-даних, що будуть доступні не тільки досвідченим розробникам, а й власникам веб-ресурсів без практичних навичок в галузі програмування.

Використання ПОМ замість мов загального призначення істотно підвищує рівень абстрактності коду, що дозволяє вести розробку швидко і ефективно, створювати легкі в розумінні та супроводі програми.

Для виявлення переліку необхідних функцій предметно орієнтованої мови проведено низку експериментів з видобування та збереження інформації з веб-ресурсів. Для видобування даних розроблено програми-парсери на мові програмування Perl.

Далі представлено два варіанти програмної реалізації однієї задачі засобами мови програмування загального призначення Perl та розробленої предметно-орієнтованої мови.

```
#!/usr/bin/perl -w
use 5.10.0;
use strict;
use Data::Dumper;
```

```
use Mojo::UserAgent;
use DBI;

my $dsn=»DBI:SQLite:dbname=store.sqlite»;
my $table=»сигналізації»;
my $url=»http://bezpeka.top/signalizatsii-gsm»;

#отримуємо вміст HTML сторінки
my $dbh = DBI->connect($dsn,undef,undef, { RaiseError => 1 }) or die $DBI::errstr;
my $ua = Mojo::UserAgent->new;
my $res=$ua->get($url)->result;
if ($res->is_error) {
 say $res->message;
 exit(1);
}

#парсинг HTML-коду
my $dom = Mojo::DOM->new();
$dom=$dom->parse($res->body);
unless ($dom) {
 say «операція парсингу HTML-коду не виконана»;
 exit(1);
}

#пошук необхідних даних та збереження їх в масив @data
my @data;
my@items=$dom->find('div[class~=»product-layout»]')->each;
unless (@items) {
 say «Список товарів не знайдено»;
 exit(1);
}
foreach my $i (@items) {
 my $link=$i->at('div.caption > h4 > a');
 my $url=$link->attr('href');
 my $name=$link->text;
 my $desc=$i->at('div.caption > p')->text;
   push @data, {url=>$url,name=>$name, desc=>$desc};
}

#створення таблиці з результатами
$dbh->do(«create table if not exists '$table' (id int primary key,name text, desc text, url text)») or die $DBI::errstr;

#збереження результатів в таблицю
my $sth=$dbh->prepare(«insert into '$table'(name,desc,url) values(?,?,?)») or die $DBI::errstr;
foreach my $e (@data) {
    $sth->execute($e->{name},$e->{desc},$e->{url}) or die $DBI::errstr;
}
```

Незважаючи на тривіальну задачу і використання «доброзичливих» модулів, які приховують більшу частину деталей, отриманий програмний код є важким для розуміння недосвідчених користувачів.

Крім того дана програма має такі недоліки:

- нестійкість до модифікацій API використаних бібліотек;
- низька репрезентативність коду, що ускладнює розуміння структури програми, спираючись на програмний код;
- незручна в налагодженні та тестуванні;
- потребує значних модифікацій в разі використання інших модулів.

Програмна реалізація на ПОМ є значно виразнішою.

```
URL http://bezpeka.top/signalizatsii-gsm
SECTION d {div[class~=»product-layout»]}
 d.title=TEXT {div.caption > h4 > a}
 d.link=LINK {div.caption > h4 > a}
 d.description=TEXT {div.caption > p}
STORE type=DB name=»сигналізації» data=d
```

Більшу частину «коду» становить інформація про структуру представлення даних на вебресурсі, яку користувач легко може отримати використовуючи вбудовані функції сучасних браузерів. Такий код легко писати, читати і модифікувати звичайним користувачам. Його не потрібно змінювати у разі використання різних методів збереження даних або використання інших алгоритмів для отримання та парсингу HTML-коду. Завдання обробки помилок, налагодження і тестування бере на себе система виконання.

Загальну схему взаємодії модулів розробленої мови представлено на рис. 2.

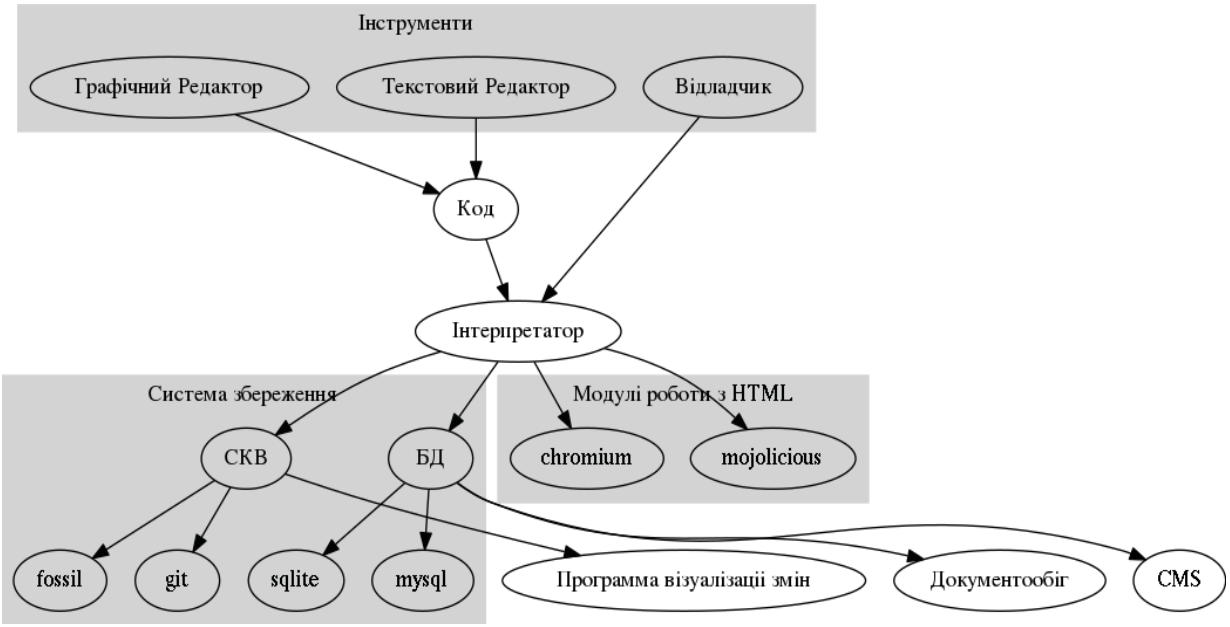

Рис. 2. Схема взаємодії компонентів предметно орієнтованої мови оперування веб-даними

Розроблена мова є декларативною, що в сукупності з візуальним редактором дає можливість користуватись нею навіть недосвідченим користувачам. Для професійних програмістів розроблено текстовий редактор. Мова підтримує два варіанти збереження даних — це бази даних та системи контролю версій, що дозволяє користувачам переглядати історію змін необхідної інформації і проводити додаткові дослідження.

Семантичний профіль відображення вмісту корпоративних веб-ресурсів. Семантичний профіль являє собою смислову розмітку всіх елементів веб-сторінки додатковими тегами та атрибутами, які дозволяють визначити не тільки стильове оформлення, але й сенс певного блоку веб-сторінки [8]. Використовуючи той же набір тегів, що і для семантичної розмітки, користувач може створювати шаблони відображення веб-сторінок певного типу в тому вигляді, який дозволить йому найбільш швидко і ефективно отримувати необхідну інформацію. При цьому він може відключити відображення тих блоків, які, на його думку, позбавлені корисної інформації.

Як приклад розглянемо структуру веб-ресурсу mebel-art.com.ua, який розміщений на першій сторінці пошукової видачі Google за запитом «меблі на замовлення».

Головна сторінка веб-ресурсу представлена на рис. 1. При стандартній роздільній здатності екрану 1920 × 1080 точок, висота «корисної» області при роботі з веб-браузером становить близько 940 точок. Сукупна довжина сторінки становить 6568 точок. А значить для перегляду всієї інформації на сторінці, користувач змушений прокрутити вниз більш шести екранів.

При цьому вся корисна інформація може бути розміщена не більше ніж на двох екранах. Інші елементи веб-сторінки створені для підвищення маркетингової привабливості сторінки або для кращої індексації сторінок пошуковими системами.

Структуру головної сторінки веб-ресурсу mebel-art.com.ua показано на рис. 3.

1 —Шапка сайту, де міститься вся необхідна користувачеві інформація: логотип, контактні дані, навігаційне меню, кошик, пошук, реєстрація. Цей блок є інформативним.

2 —Блок «заклик до дії (call to action)», в якому розміщена графічна ілюстрація і форма виклику замірника. Призначення таких блоків суто маркетингове, спонукати відвідувача до

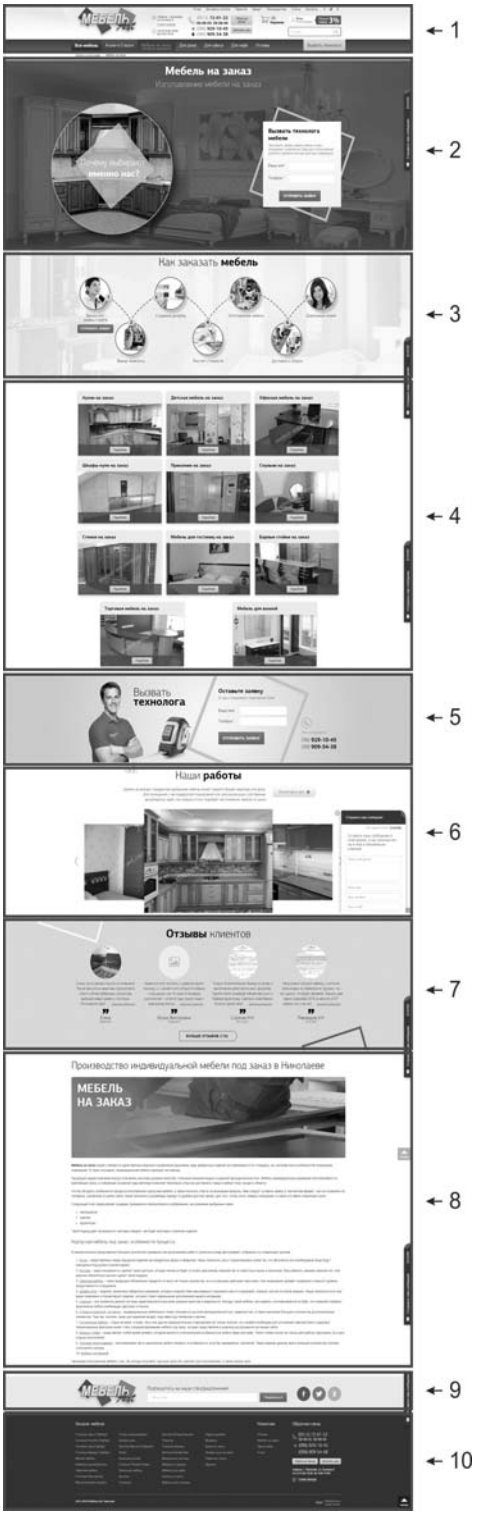

Рис. 3. Структура головної сторінки веб-ресурсу mebel-art.com.ua

негайних дій, викликати майстра, замовити зворотний дзвінок, забронювати столик тощо. Неінформативний, оскільки виклик замірника, як правило, здійснюється тільки після дзвінка і узгодження всіх деталей.

3 —Блок «схема роботи» також ємаркетинговим, який візуалізує схему замовлення, але не надає користувачеві ніякої конкретної інформації про деталі роботи компанії — терміни, вартість, варіанти оплати і т.ін. Неінформативний, тому що не скасовує необхідності здійснювати дзвінок і з'ясовувати деталі.

4 —Блок «проекти» є інформативним, тому що дає користувачеві можливість ознайомитися з продукцією, яку надає фірма; у цьому блоці представлено саме ту інформацію, яка необхідна користувачеві для прийняття рішення.

5 —Блок «заклик до дії», аналогічний блоку №2 — неінформативний.

6 —Блок «наші роботи», який по суті дублює вміст блоку №4, але, на відміну від нього, тут інформація представлена в незручному вигляді, без рубрикації — неінформативний.

7 —Блок «відгуки клієнтів» дозволяє ознайомитися з реальними відгуками клієнтів; інформативний, тому що допомагає користувачеві прийняти рішення.

8 —Блок «стаття» містить зображення і текст. Не містить корисної для прийняття рішення інформації, але є необхідною умовою для успішного просування веб-ресурсу в пошуковій видачі. Неінформативний, тому що створений спеціально для індексації пошуковою машиною.

9 —Блок «заклик до дії» неінформативний.

10 —Нижня частина сайту, в якій дублюється представлена в шапці контактна інформація. Посилання на сторінки каталогу дублюють навігаційне меню в шапці і сприяють кращій індексації сайту пошуковими системами «в глибину». Неінформативний, тому що створений спеціально для індексації пошуковою машиною.

Аналіз веб-ресурсу показав, що інформативними є блоки 1, 4, 7; висота їх становить 1736 точок. Тобто інформативний (корисний для користувача) вміст сторінки становить 26, 4 відсотків. На рис. 4 показано зовнішній вигляд

Рис. 4. Веб-сторінка з відключеним відображенням неінформативних блоків

веб-ресурсу з відключеним відображенням неінформативних блоків.

Використовуючи також можливості нової *DSL*-мови був створений шаблон відображення для сайтів компаній, який дозволяє відображати подану інформацію більш компактно. Результати роботи подано на рис. 5.

Важливо відзначити, що шаблон можна застосовувати автоматично до всіх веб-ресурсів даного виду. Тобто всі веб-ресурси компаній будуть відображатися з тим же стильовим оформленням (колір, розміри блоків, шрифти, розмір тексту тощо)

Для підвищення ефективності роботи з новою мовою розроблений візуальний інтерфейс, заснований на технології *WYSIWYG*. Для розмітки існуючих веб-ресурсів розроблено візуальний інтерфейс, в якому відобра-

Рис.5. Веб-сторінка після застосування шаблону відображення

жається веб-сторінка повністю. При натисканні на будь-який текстовий або графічний елемент сторінки підсвічується елемент розмітки *DOM* і з'являється вікно для введення тега розмітки вручну або для вибору зі списку готових тегів.

Для розробки нових веб-ресурсів за стандартами семантичної розмітки розроблено систему керування вмістом веб-ресурсу (*CMS*), в якій теги семантичної розмітки вбудовані в ядро системи і автоматично інтегруються в код на етапі генерації веб-сторінки.

Можливості мови дозволяють користувачам створювати шаблони відображення веб-сторінок на основі семантичного профілю веб-ресурсу.

З цією метою використовується візуальний інтерфейс, в якому користувач може шляхом перетягування мишею елементів створювати шаблон відображення для певного типу веб-ресурсів (сайт компанії, інтернет-магазин, новинний сайт, блог).

Семантична розмітка веб-ресурсів у зв'язці з універсальним граббером може принципово змінити порядок взаємодії користувача з пошуковою системою.

Поняття веб-ресурсу зміниться і більше не буде самостійним сайтом зі своїм строго визначеним дизайном, який відображається по зверненню до доменного імені. Новий підхід визначає веб-ресурс як набір даних і семантичний профіль, складені за певними правилами.

На підставі семантичного профілю дані можуть бути відображені в будь-якому зручному для користувача вигляді на підставі його особистого шаблону відображення для відповідного типу веб-ресурсів.Користувач отримує можливість довільно змінювати структуру веб-ресурсу, обирати, які саме елементи вебсторінки будуть відображатися, а які будуть проігноровані.

Також семантичний профіль дозволяє оперувати даними поза доменним іменем, на-

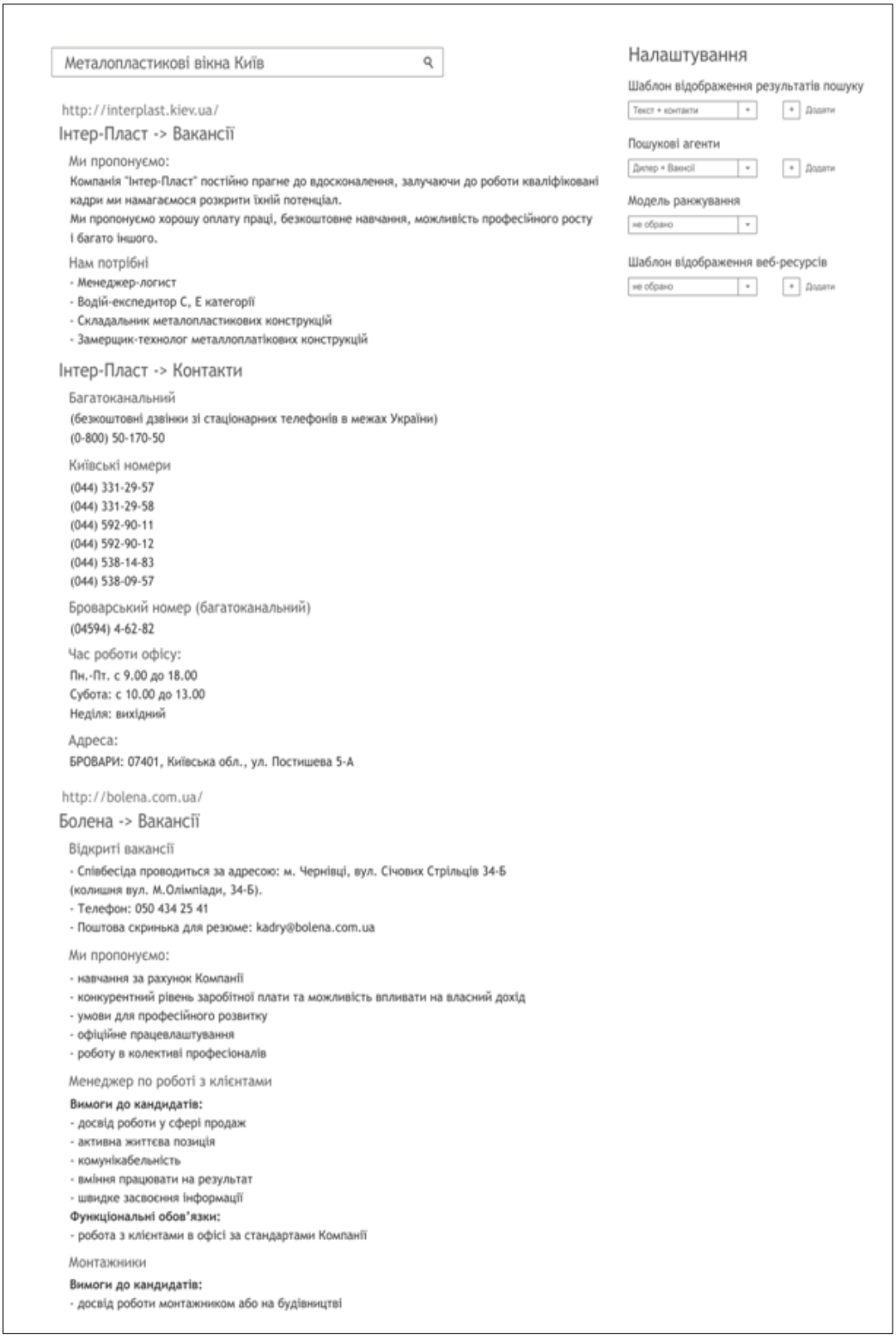

Рис. 6. Результати обробки пошукових даних

приклад, порівнювати прямо на етапі пошуку певні види послуг або товарів, застосовувати фільтри і сортування.

Модуль пошуку інформації. Модуль інформаційного пошуку реалізовано у вигляді метапошукової системи, що як базові результати використовує пошукову видачу *Google.com.ua*.

В системі реалізовано нові можливості, для підвищення якості та зручності пошуку інформації для користувача:

- вибір альтернативної моделі ранжування результатів пошуку на основі оцінок користувачів;
- вибір шаблонів відображення результатів пошуку;
- розробка власних шаблонів відображення результатів пошуку;
- застосування пошукових агентів для постобробки пошукової видачі;
- розробка власних пошукових агентів

Результати пошуку за запитом «металопластикові вікна Київ» із застосуванням шаблону «Вакансії+контакти», що відображає повний текст сторінок вакансії та контакти, а також застосуванням пошукового агенту, який відбирає в пошуковій видачі *Google* всі компанії, що є офіційними дилерами або виробниками і в яких є відкриті вакансії подано на рис. 6.

Спочатку результати пошуку *Google* додатково обробляються пошуковим агентом, зберігаються в базу даних, а потім виводяться на екран згідно шаблону відображення.

В лівій колонці відображаються поле для вводу пошукового запиту та результати пошуку. В правій колонці розташовані налаштування пошуку:

- вибір шаблону відображення результатів пошуку – надає можливість обрати, яка саме інформація з веб-сторінки і в якому вигляді буде відображена на сторінці результатів. Реалізовано також можливість додавання власних шаблонів;
- пошукові агенти – дозволяє обрати один з існуючих пошукових агентів для більш детального аналізу пошукової видачі. Також реалізовано можливість додавання власних пошукових агентів;
- вибір моделі ранжування;
- шаблон відображення веб-ресурсів – дозволяє обрати шаблон для відображення вмісту конкретного веб-ресурсу при переході за посиланням з пошукової видачі.

Висновки

За використання програм-парсерів, реалізованих засобами мови програмування *Perl*, досліджено процес видобування даних з веб-ресурсів. На основі отриманих даних було виявлено перелік необхідних функцій, використаних при розробці нової предметно орієнтованої мови оперування вебданими

За допомогою автоматичного парсера, реалізованого засобами представленої в статті мови проаналізовано інформаційний вміст 1000 корпоративних веб-ресурсів. На базі отриманих даних побудовано загальну структуру корпоративних сайтів.

Дослідження існуючих словників семантичної розмітки показало, що як базовий набір класів для опису семантичної структури корпоративних веб-ресурсів доцільно використовувати стандарти *Good Relations*, який є спеціалізованим розширенням найбільш поширеного словника Schema.org для електронної комерції. Враховуючи базову онтологію, було додано нові класи згідно структури корпоративних веб-ресурсів.

На базі створеної ПОМ та словника семантичної розмітки корпоративних вебресурсів, розроблено програмний комплекс, що реалізує роботу всіх модулів описаної системи КОДІ.

Розроблена предметно орієнтована мова для оперування веб-даними є ефективним інструментом для збору, зберігання і відображення вмісту веб-ресурсів. Використання декларативного підходу в сукупності з візуальним редактором дає можливість користуватись нею навіть недосвідченим користувачам.Також розроблена ПОМ є ефективною платформою для створення пошукових агентів на основі семантичної розмітки.

Використання семантичного профілю веб-ресурсу дозволяє відображати інформацію в будь-якому зручному для користувача вигляді на підставі його особистого шаблону відображення для відповідного типу веб-ресурсів. Користувач отримує можливість довільно змінювати структуру веб-ресурсу.

Модуль інформаційного пошуку, реалізований у вигляді метапошукової системи, значно підвищує ефективність пошуку за використання шаблонів відображення результатів пошуку та пошукових агентів.

Розроблена система надає користувачу ефективні, зручні та прості у використанні інструменти обробки веб-даних на всіх етапах взаємодії з семантичною павутиною від створення веб-ресурсів до пошуку інформації.

СПИСОК ЛІТЕРАТУРИ

1. Zosimov V., Stepashko V., Bulgakova O.Inductive Building of Search Results Ranking Models to Enhance the Relevance of Text Information Retrieval. "Database and Expert Systems Applications, Valencia, Spain / Ed. by Markus Spies at al. — Los Alamitos: IEEE Computer Society, 2015. — 316 p. / — P. 291—295
2. Zosimov V., Stepashko V., Bulgakova O. Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results:. — Artificial Intelligence Methods and Techniques for Business and Engineering Applications — Rzeszow, Poland; Sofia, Bulgaria: ITHEA, 2012. — 99—112 pp.
3. Zosimov V., Bulgakova O. Usage of Inductive Algorithms for Building a Search Results Ranking Model Based on Visitor Rating Evaluations. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018, IEEE,September, Pages 2018 466—470
4. Zosimov V. Prospects for Applying the Concept of the Semantic Web Analysis for Existing sites. Індуктивне моделювання складних систем, : Зб. наук. пр. — К.: МННЦ ІТС НАН та МОН України, 2014. — Вип. 6. — С. 41—46.
5. Van Deursen A., Klint P., Visser J. Domain-Specific Languages: An Annotated Bibliography. ACM SIGPLAN Notices. Vol. 35, N 6.2000. P. 26—36.
6. Сухов А.О. Сравнение систем разработки визуальных предметно-ориентированных языков. Математика программных систем: межвузовский сборник научных статей / Перм. гос. нац. исслед. ун-т. — Пермь, 2012. — С. 84—111.
7. Казакова А.С. Методы и инструменты реализации предметно-ориентированных языков программирования. Системное программирование. 2009. Т. 4. С. 51—80.
8. Zosimov V. Khrystodorov O., Bulgakova O. Dynamically changing user interfaces: software solutions based on automatically collected user information. Proceedings of the Institute for System Programming, vol 30:3 3, 2018, P. 207—220. DOI: 10.15514/ISP-2018-30(3)-15

Надійшла 21.11.2018

REFERENCES

1. Zosimov V., Stepashko V., Bulgakova O. 2015. Inductive Building of Search Results Ranking Models to Enhance the Relevance of Text Information Retrieval. "Database and Expert Systems Applications, Valencia, Spain. Ed. by Markus Spies at al. — Los Alamitos: IEEE Computer Society. pp. 291—295
2. Zosimov V., Stepashko V., Bulgakova O. 2012. Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results:. — Artificial Intelligence Methods and Techniques for Business and Engineering Applications. Rzeszow, Poland; Sofia, Bulgaria: ITHEA. pp. 99—112.
3. Zosimov V., Bulgakova O. 2018. Usage of Inductive Algorithms for Building a Search Results Ranking Model Based on Visitor Rating Evaluations. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018, IEEE. pp. 466—470
4. Zosimov V. 2014. Prospects for Applying the Concept of the Semantic Web Analysis for Existing sites. Inductive Modelling of complex systems : K.: ISSC ITSNACandMOCUkraine, 6. C. 41—46.
5. Van Deursen A., Klint P., Visser J. 2000. Domain-Specific Languages: An Annotated Bibliography. ACM SIGPLAN Notices. Vol. 35, N 6. pp. 26—36.
6. SuchovA.O. 2012. Comparison of visual object-oriented languagesdevelopment systems. Mathematics of software systems: intercollegiate collection of scientific articles . Perm. gov. nat. research. un-t. Perm. pp. 84—111.

7. KazakovaA.S. 2009. Methods and tools for implementing domain-specific programming languages. System programming. Vol. 4. pp. 51—80.
8. Zosimov V. Khrystodorov O., Bulgakova O. 2018. Dynamically changing user interfaces: software solutions based on automatically collected user information. Proceedings of the Institute for System Programming, vol 30:3 3. pp. 207—220. DOI: 10.15514/ISP-2018-30(3)-15

Received 21.11.2018

Zosimov V.V., Ph.D in Techn.Sciences,
Associate Professor of the Computer Science and Applied Mathematics Department
V.O. Sukhomlynsky Mykolaiv National University,
Nikolska str., 24, Mykolaiv, 54030, Ukraine
zosimovvv@gmail.com

TECHNOLOGY OF WEB APPLICATIONS BASED ON THE CYBER-ENTITIES IDENTIFICATION

Introduction.The rapid development of information technology in recent decades has put for the society the tasks of effectively process large volumes of poorly structured information presented in the form of web pages. Among them, the standards research for the certain types of web resources development, the search, extraction, processing, analysis, storage and display of information.

Purpose. Development of the integrated web-data operating system within the concept of the semantic web, focused on improving the methods of information search, the creation of web resources with integrated semantic markup and programme search agents.

Methods. Methods of analysis and data processing, Data mining, Web Mining, machine learning methods, group method of data handling, modern methods for constructing software products with modular architecture, search agents developing methods, semantic markup inte-grating methods.

Results. Based on the research results, a general structure and semantic markup dictionary for corporate web resources, the domain specific language of web data operating and the soft-ware package that implements the work of all modules of the described system for the web-data complex operating are developed.

Conclusions. The proposed system of integrated web-data operating, provides the user with efficient, convenient and easy to use tools for processing web data at all stages of interaction with the world wide web from the web resources creation to the information search. The system is based on a new domain specific web-data operating language and the corporate web resources semantic markup dictionary.

Keywords: *information search, ranking model, meta-search system, semantic markup, ontology, web resource structure, Data mining, domain specific language, content management system.*

Зосимов В.В., кандидат технических наук, доцент,
Николаевский национальный университет им. В.А. Сухомлинского,
Николаев, Украина,
zosimovvv@gmail.com

КОМПЛЕКСНЫЙ ПОДХОД К ПОВЫШЕНИЮ ЭФФЕКТИВНОСТИ ОБРАБОТКИ ВЕБ-ДАННЫХ НА ОСНОВЕ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ

Введение. Стремительное развитие информационных технологий в последние десятилетия поставило перед обществом целый ряд задач по эффективной обработке больших объемов слабоструктурированной информации, представленной в виде веб-страниц. Среди них — исследование стандартов разработки определенных видов веб-ресурсов, поиск, извлечение, обработка, анализ, хранение и отображение информации.

Цель. Разработка системы комплексного оперирования веб-данных в рамках концепции семантической паутины, направленной на совершенствование методов поиска информации, создание веб-ресурсов с интегрированной семантической разметкой и программных поисковых агентов.

Методы. Методы анализа и обработки данных, DataMining, WebMining, методы машинного обучения, метод группового учета аргументов, современные методы построения программных продуктов с модульной архитектурой, методы разработки поисковых агентов и интеграции семантической разметки.

Результаты. На основе результатов исследования были разработаны: общая структура и словарь семантической разметки для корпоративных веб-ресурсов, предметно-ориентированный язык обработки веб-данных и программный комплекс, реализующий работу всех модулей описанной системы комплексной обработки веб-данных.

Выводы. Представленная система комплексной обработки веб-данных предоставляет пользователю эффективные, удобные и простые в использовании инструменты для обработки веб-данных на всех этапах взаимодействия со всемирной паутиной — от создания веб-ресурсов до поиска информации. В основу системы положен новый предметно-ориентированный язык обработки веб-данных и словарь семантической разметки корпоративных веб-ресурсов.

Ключевый слова: *поиск информации, модель ранжирования, метапоисковая система, семантичная розметка, онтологии, структура веб-ресурса, извлечение знаний из данных, предметно-ориентированный язык, система управления содержанием*