Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

ЗОСІМОВ В'ячеслав Валерійович

УДК 004.9

МОДЕЛІ ТА ЗАСОБИ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ДАНИХ КОРПОРАТИВНИХ ВЕБ-РЕСУРСІВ

01.05.03 – математичне та програмне забезпечення
обчислювальних машин і систем

Автореферат
дисертації на здобуття наукового ступеня
доктора технічних наук

Київ – 2019

Дисертацією є рукопис.

Роботу виконано на кафедрі теорії та технології програмування факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка, МОН України.

Науковий консультант: член-кореспондент НАН України, доктор фізико-математичних наук, професор
Анісімов Анатолій Васильович,
Київський національний університет імені Тараса Шевченка,
декан факультету комп'ютерних наук та кібернетики

Офіційні опоненти: доктор технічних наук, професор
Литвиненко Володимир Іванович
Херсонський національний технічний університет,
завідувач кафедри інформатики і комп'ютерних наук

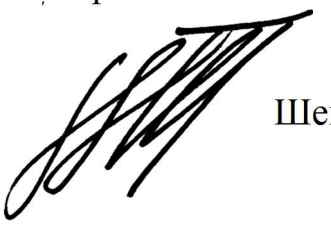
доктор технічних наук, доцент
Осипенко Володимир Васильович
Київський національний університет технологій та дизайну,
професор кафедри комп'ютерної інженерії та електромеханіки

доктор технічних наук, доцент
Глибовець Андрій Миколайович
Національний університет «Києво-Могилянська академія»,
доцент кафедри мережних технологій

Захист відбудеться 03 жовтня 2019 р. о 14 год. на засіданні спеціалізованої вченої ради Д 26.001.09 Київського національного університету імені Тараса Шевченка за адресою: м. Київ, проспект академіка Глушкова, 4-Д, факультет комп'ютерних наук та кібернетики, ауд. 01.

З дисертацією можна ознайомитися у Науковій бібліотеці ім. М. Максимовича Київського національного університету імені Тараса Шевченка за адресою: 01601, м. Київ, вул. Володимирська, 58, зал №12.

Автореферат розісланий «____» серпня 2019 р.

Вчений секретар спеціалізованої вченої ради  Шевченко В.П.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Стрімкий розвиток інформаційних технологій в останні десятиріччя ставить перед суспільством цілий клас задач ефективної обробки великих масивів слабко структурованої інформації, представленої у вигляді веб-сторінок. Серед них, створення стандартів розробки веб-ресурсів визначених типів, пошук, видобування, обробка, аналіз, збереження та відображення інформації.

Експоненціальне зростання кількості інформаційних джерел зумовлює розвиток таких напрямків, як інтелектуальний аналіз даних (Data mining), видобування даних з веб-сторінок (Web mining), методи інтелектуального пошуку інформації, машинне навчання, кластерний та регресійний аналіз.

Пошук та аналіз текстової інформації досліджували видатні вітчизняні та зарубіжні вчені (Анісімов А. В., Маккарті Д., Глибовець М.М., Ньюелл А., Марченко О.О., Люгер Д. Ф., Фостер Д. М., Поспєлов Д. А., Рубашкін В. Ш., Виноград Т. В., Ланде Д. В., Попов Е. В., Глибовець А.М., Піттс У., Осуга С., Лакофф Дж., Хопкрофт Дж., Леонтьєв Н. Н., Шемакін Ю. І., Гальперін І. Р., Кобозева І. М., Мельчук І. А., Кронгауз М. А., Греймас А. Ж., Севбо І.П., Широков В. А.). Та все ж відкритими залишаються питання створення методів автоматичного видобування інформації з веб-ресурсів, інтеграції семантичної розмітки в HTML-код, альтернативних методів відображення вмісту веб-ресурсів без прив’язки до URL-адреси веб-сторінки, персоналізації результатів пошуку.

Розробка концепції семантичної павутини стала наступним кроком розвитку глобальної мережі. Розміщена в мережі Інтернет інформація зручна для розуміння людиною. В свою чергу, 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 у веб-середовищі.

Поява стандартів семантичної розмітки відкриває широкі перспективи для використання інструментів інтелектуального пошуку інформації в мережі Інтернет – пошукових агентів. За класифікацією Стефена Хаага це агенти по видобуванню та аналізу даних, які працюють в сховищах даних. При розробці метапошукових систем, пошукові агенти використовуються для додаткової обробки та аналізу результатів роботи пошукової системи. Цей напрямок розвивається досить повільно через відсутність інструментів розробки пошукових агентів з відкритим початковим кодом, доступних широкому колу користувачів.

Враховуючи все вищезазначене, актуальною є задача розробки комплексної системи оперування веб-даними, що буде включати в себе методи, моделі та програмні засоби оперування веб-даними, зокрема, представленими на корпоративних веб-ресурсах вітчизняного сегменту всесвітньої павутини, на семантичному рівні в режимі реального часу, в умовах слабкої структурованості, неоднорідності та неповноти інформації, а також постійно зростаючої кількості джерел.

Зв'язок роботи з науковими програмами, планами, темами. Основні дослідження за темою дисертації здійснено згідно з планами наукових досліджень кафедри комп'ютерних наук та прикладної математики механіко-математичного факультету Миколаївського національного університету ім. В.О. Сухомлинського, в тому числі в рамках науково-дослідних робіт: «Моделювання соціально-економічних явищ та екологічних процесів південно-західного регіону України на основі індуктивних моделей» (реєстраційний номер в УкрІНТЕІ 0111U010622, 2012-2013 рр.); «Моделі та методи інтелектуального аналізу даних в предметно-орієнтованій інформаційній системі» (реєстраційний номер в УкрІНТЕІ 0115U001249, 2015-2017 рр.), кафедри теорії та технології програмування факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка при виконанні фундаментальної теми «Теорія і методи розробки інтелектуальних інформаційних технологій та систем» (№ 0116U006378, квітень 2016 – грудень 2020 рр.).

Мета і завдання дослідження. Метою дослідження є підвищення якості інтелектуального пошуку та обробки даних в мережі Інтернет із застосуванням програмних пошукових агентів, завдяки розробці нових методів, моделей та засобів оперування веб-даними, зокрема, представленими на корпоративних веб-ресурсах вітчизняного сегменту всесвітньої павутини, на семантичному рівні в режимі реального часу в умовах слабкої структурованості, неоднорідності та неповноти інформації, а також постійно зростаючої кількості джерел.

Для досягнення основної мети необхідно виконати наступні завдання:

- провести аналіз інформаційного вмісту та елементів навігації корпоративних веб-ресурсів для побудови їх загальної структурної схеми. На її основі створити семантичний профіль корпоративного веб-ресурсу вітчизняного сегменту всесвітньої павутини, що представляє собою онтологію семантичних класів, яка базується на загальновживаній в світовій практиці глобальній онтології для опису інформаційного вмісту всіх типів веб-ресурсів, Schema.org;
- провести аналіз найбільш затребуваних функцій для оперування веб-даними, які використовуються в процесі видобування, обробки, збереження та відображення даних з веб-ресурсів. Ґрунтуючись на отриманих результатах створити предметно-орієнтовану мову оперування веб-даними, основним призначенням якої є реалізація всіх стандартних функцій оперування даними у вигляді вбудованих засобів мови;
- дослідити методи розробки веб-ресурсів, а також методи інтеграції семантичної розмітки в HTML-код вже існуючих та нових веб-сторінок. На основі отриманих даних розробити систему керування вмістом корпоративних веб-ресурсів з вбудованою можливістю інтеграції семантичної розмітки безпосередньо в процесі генерації HTML-коду веб-сторінки, для скорочення часу розробки веб-ресурсу та актуалізації змін інформації в реальному часі;
- використовуючи вбудовані засоби оперування веб-даними розробленої предметно-орієнтованої мови, створити метод автоматизованої інтеграції

семантичної розмітки в HTML-код вже існуючих веб-сторінок, що не потребує редагування початкового коду веб-ресурсу;

- ґрунтуючись на результатах побудови семантичного профілю корпоративних веб-ресурсів, створити метод відображення інформаційного вмісту на основі користувацьких шаблонів, що забезпечить відображення всіх веб-ресурсів певного типу згідно однієї структури розташування блоків та стильового оформлення, а також відсів даних, нерелевантних інформаційним потребам користувача;

- розробити метод підвищення ефективності пошуку інформації в мережі Інтернет по ключовим словам, заснований на додатковій обробці результатів роботи сучасних пошукових систем;

- дослідити шляхи створення моделі ранжування результатів пошуку інформації, заснованої на суб'єктивних для кожного користувача показниках. Розробити методику її побудови, що надасть можливість персоналізувати результати пошуку, зробити їх менш залежними від факторів, на які можуть впливати власники веб-ресурсів. Використати експертне навчання для покращення результатів ранжування;

- об'єднати всі запропоновані методи, моделі та засоби в систему комплексного оперування даними в мережі Інтернет та розробити програмний комплекс, що реалізує її функціональні можливості.

Об'єктом дослідження є оперування даними, представленими в мережі Інтернет, зокрема видобування, збереження, обробка, відображення інформації, представленої на веб-сторінках у вигляді HTML-коду, а також інтелектуальний пошук та аналіз веб-даних.

Предметом дослідження є методи видобування даних з веб-ресурсів без прив'язки до структури вихідного коду, моделі та методи персоналізації ранжування результатів пошуку інформації, моделі та методи створення предметно-орієнтованих мов обробки даних, методи побудови спеціалізованих онтологій для певних типів веб-ресурсів та інтеграції семантичної розмітки в HTML-код веб-сторінок, методи створення гнучких архітектурних рішень для побудови ефективних програмних систем.

Методи дослідження. Методи аналізу та інтелектуальної обробки даних Data Mining, зокрема, методи видобування інформації з веб-сторінок Web Mining, методи машинного навчання, метод групового урахування аргументів, сучасні методи побудови програмних продуктів з модульною архітектурою, методи розробки пошукових агентів, методи інтеграції семантичної розмітки в HTML-код.

Наукова новизна отриманих результатів

Наукова новизна отриманих результатів полягає в теоретичному обґрунтуванні та практичній реалізації комплексного підходу для вирішення проблеми підвищення якості інтелектуального пошуку та обробки даних в мережі Інтернет із застосуванням програмних пошукових агентів, що об'єднує нові методи, моделі та засоби оперування веб-даними на семантичному рівні в режимі реального часу.

Отримано нові результати:

Вперше на основі аналізу інформаційного вмісту та елементів навігації 1000 корпоративних веб-ресурсів побудовано їх загальну структуру та семантичний профіль, що являє собою набір семантичних класів та властивостей для опису їх інформаційного наповнення.

Розроблено нову предметно-орієнтовану мову оперування веб-даними з широким функціоналом для видобування, обробки, збереження та відображення веб-даних, яка відрізняється тим, що всі можливості обробки інформації реалізовані в якості вбудованих засобів мови і орієнтовані в першу чергу на роботу з об'єктною моделлю документів, представлених у всесвітній павутині у вигляді HTML-коду. Основною перевагою даної мови є те, що вона реалізована в декларативному стилі і дозволяє розробнику описувати складні операції з даними простими вбудованими функціями. Всі можливості мови реалізовано з високим рівнем абстракції, що дозволяє змінювати окремі внутрішні механізми обробки даних, таких як формат збереження даних, методи розбору структури веб-сторінок, без необхідності корегування програмного коду. Розроблена мова стала ядром системи підвищення ефективності обробки веб-даних, та базою для реалізації платформи створення пошукових агентів, заснованої на семантичній розмітці веб-ресурсів.

Запропоновано новий підхід до унікальної ідентифікації фрагментів даних веб-сторінок в рамках доменного імені, який відрізняється тим, що заснований на використанні семантичних атрибутів HTML-блоків замість URL-адреси батьківської веб-сторінки. Застосування даного підходу дозволяє формувати шаблони відображення інформаційного вмісту веб-ресурсу з довільним розташуванням блоків даних, не зважаючи на їх приналежність до певних сторінок веб-ресурсу.

Розроблено новий метод відображення інформаційного вмісту веб-ресурсів на основі користувацьких шаблонів. Він відрізняється принципово новим підходом, що базуючись на онтології корпоративних веб-ресурсів та унікальній семантичній ідентифікації фрагментів даних, передбачає створення на стороні користувача одного шаблону з визначеною структурою розташування блоків даних та їх стильовим оформленням, для відображення всіх веб-ресурсів певного типу. Метод користувацьких шаблонів може бути ефективно застосований не тільки для відображення вмісту веб-ресурсу, а й для відображення результатів пошукової видачі. Оскільки кожна одиниця інформації має унікальний семантичний ідентифікатор, це дає змогу сформувати шаблон відображення результатів пошукової видачі в довільному вигляді.

Розроблено нову персоналізовану модель ранжування результатів Інтернет-пошуку на основі оцінок користувачів. Модель ранжування відрізняється тим, що заснована на формуванні унікальних експертних груп для кожного користувача. Експерти відбираються на основі показника ваги відносно поточного користувача. Вага розраховується як міра узгодженості думок поточного користувача з потенційними експертами і використовується в якості коефіцієнта при розрахунку рангу веб-ресурсу на основі експертних оцінок.

Вперше розроблено абстрактну модель системи комплексного оперування даними в мережі Інтернет, яка об’єднує всі представлені в роботі підходи, методи, моделі та засоби.

Одержало подальший розвиток:

Концепція «Семантичної павутини», а саме метод інтеграції семантичної розмітки до вже існуючих веб-ресурсів, який відрізняється створенням додаткового семантичного профілю веб-ресурсу, що містить семантичний опис всіх блоків даних, зберігається в базі даних та накладається на вихідний HTML-код веб-сторінки в процесі його генерації, метод розробки веб-ресурсів з інтегрованою семантичною розміткою, який відрізняються додаванням семантичної розмітки безпосередньо в процесі генерації HTML-коду веб-сторінки в реальному часі, що значно спрощує подальшу актуалізацію даних, аналіз вмісту веб-ресурсів, видобування даних, їх обробку, та розширює можливості застосування методів інтелектуального пошуку.

Метод побудови пошукових агентів, який відрізняється способом унікальної ідентифікації фрагментів даних в межах доменного імені на основі семантичних атрибутів, а також застосуванням засобів нової предметно-орієнтованої мови для видобування, обробки та зберігання веб-даних.

Удосконалено:

Технологію пошуку інформації в мережі Інтернет по ключових словах, що відрізняється новим підходом до відображення результатів пошуку, новою персоналізованою моделлю ранжування результатів, а також можливістю застосування пошукових агентів безпосередньо в процесі пошуку.

Програмно реалізовано:

- модульну систему комплексного оперування даними в мережі Інтернет на основі розробленої абстрактної моделі;
- предметно-орієнтовану мову оперування веб-даними;
- метапошукову систему, засновану на додатковій обробці результатів роботи сучасних пошукових систем із застосуванням користувацьких шаблонів відображення результатів пошуку, нової моделі ранжування на основі оцінок користувачів, а також пошукових агентів;
- платформу розробки пошукових агентів на базі розробленої предметно орієнтованої мови оперування веб-даним та семантичного профілю корпоративних веб-ресурсів;
- платформу розробки шаблонів відображення вмісту веб-ресурсів на основі особистих потреб користувача;
- систему керування вмістом веб-ресурсів для створення веб-ресурсів з інтегрованою семантичною розміткою;
- інтерфейс інтеграції семантичної розмітки в HTML-код вже існуючих веб-сторінок.

Практичне значення отриманих результатів. Розроблені в роботі моделі, методи, підходи та засоби підвищення ефективності оперування веб-даними, практично реалізовані у вигляді програмного комплексу, що забезпечує інтернет-користувачів та веб-розробників вдосконаленими інструментами

інформаційного пошуку, додаткової обробки результатів роботи пошукових систем, інтеграції семантичної розмітки до вже існуючих та нових веб-ресурсів.

Результати дисертаційного дослідження впроваджено у науково-дослідну роботу механіко-математичного факультету Миколаївського національного університету ім. В.О. Сухомлинського (МНУ), та науково-дослідну роботу кафедри інформатики і комп'ютерних наук Херсонського національного технічного університету (ХНТУ), в автоматизовану систему управління навчальним процесом МНУ та ХТНУ. Розроблені в роботі методи та засоби впроваджено у виробничий процес проектної організації ТОВ «Миколаївводпроект», ТОВ «Проф-контракт», ТОВ «ЮФ Центр консультацій з питань господарської діяльності», веб-студії ТОВ «Сайт-плюс», ТОВ «Центр комп'ютерних технологій і інновацій», громадської організацій «Асоціація автомобілістів України». Також результати дисертаційного дослідження використовуються в навчальному процесі МНУ МОН України при підготовці фахівців за спеціальністю «Комп'ютерні науки» ступінь бакалавра та магістра (в теоретичній та практичній підготовці за дисциплінами «Веб-технології та веб-дизайн», «Методи та системи штучного інтелекту», «Сучасні технології веб-програмування та електронної комерції») та в навчальному процесі ХТНУ МОН України при підготовці фахівців за спеціальністю «Комп'ютерні науки» ступінь бакалавра та магістра (в теоретичній та практичній підготовці за дисциплінами «Інтелектуальний аналіз даних і знань», «Інтернет-технології опрацювання консолідованих інформаційних ресурсів», «Проектування консолідованих інформаційних ресурсів та систем», «Комп'ютерні технології та програмування»), що підтверджено актами впровадження у навчальний процес.

Особистий внесок здобувача. Робота є узагальненням результатів теоретичних та експериментальних досліджень, виконаних автором самостійно і опублікованих в особисто підготовлених працях [9, 11, 12, 13, 14, 15, 17, 19, 20, 21, 34, 36, 40, 42]

У роботах, що опубліковані у співавторстві, здобувачу належить: технологія підвищення ефективності пошуку інформації в мережі Інтернет [1], інформаційний аналіз управління веб-процесами в глобальній мережі [2], дослідження ефективності застосування ітераційних алгоритмів в задачах веб-моделювання [3], програмування розподілених процесів в задачах моделювання складних систем [4], програмна реалізація роботи ітераційного алгоритму МГУА [5, 8, 25], технологія підвищення ефективності пошуку релевантної інформації в мережі Інтернет [6], розробка моделі ранжування [7, 28], розробка методів підвищення швидкодії складних обчислень [10], розробка комп'ютерної програми віддаленого керування електронними пристроями [16], розробка алгоритму адаптації інтерфейсів веб-додатків [18], розробка програмного комплексу [22], дослідження метапошукових систем та пошукових агентів [23], методи якості інформаційного аналізу веб-ресурсів [24], методи адаптації інтерфейсів веб-додатків під потреби користувача [26, 27], технологія підвищення ефективності пошуку інформації [29], дослідження ефективності узагальненого ітераційного алгоритму [30], модель ранжування результатів

пошуку на основі оцінок користувачів [31], технологія підвищення ефективності пошуку інформації в мережі Інтернет [32], розробка програмного комплексу [33], адаптація алгоритму індуктивного моделювання для використання в GRID-системах [35], метод видобування даних з веб-сторінок [37], дослідження шляхів застосування інтелектуальних методів аналізу даних в задачах моніторингу економічних процесів [38], модель видобування знань із застосуванням індуктивних алгоритмів [39], метод ідентифікації та видобування даних на основі семантичної розмітки [41], семантична та стильова структура веб-ресурсів [43].

Апробація результатів дисертації. Основні результати роботи оприлюднені в доповідях 11 міжнародних та 1 всеукраїнській наукових конференціях серед них V International Conference on Intelligent Information and Engineering Systems INFOS 2012 (17-21 Вересня, 2012 р. м. Криниця, Польща), XIV міжнародна науково-технічна конференція «Системний аналіз та інформаційні технології» SAIT (24 квітня, 2012 р., м. Київ), Міжнародна науково-практичні конференція «Інформаційні та комунікаційні технології в науці, освіті та виробництві (29-30 жовтня, 2012 р., м. Тирасполь, Молдова), IV Міжнародний семінар з індуктивного моделювання IWIM-2012 (8-14 липня, 2012 р., м. Київ), Міжнародна конференція з індуктивного моделювання ICIM-2013 (21-24 липня, 2013 р., м. Київ), Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», ISDMCI'2013 (20-24 травня, 2013 р., м. Євпаторія), Міжнародна науково-технічна конференція «Геоінформаційні системи та комп'ютерні технології еколого-економічного моніторингу» (9-11 квітня, 2014 р., м. Дніпро), Міжнародна конференція «Комп'ютерне моделювання в наукомістких технологіях» » (28-31 травня, 2014 р., м. Харків), XVII Міжнародна конференція «Проблеми теоретичної кібернетики» (16-20 червня, 2014 р., м. Казань), Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», ISDMCI'2015 (25-28 травня, 2015 р. с. Залізний порт), XII International Workshop on Text-based Information Retrievel In conjunction with DEXA 2015, TIR 2015 (1-4 вересня, м. Валенсія, Іспанія), V Міжнародний семінар з індуктивного моделювання, IWIM-2015 (20-24 липня, 2015 р., м. Київ), Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», ISDMCI'2016 (24-28 травня, 2016 р., с. Залізний порт), XII Міжнародна конференція «Комп'ютерні науки та інформаційні технології», CSIT 2017, IWIM 2018 (05-08 вересня, 2017 р., м. Львів), Всеукраїнська наукова конференція «Інформаційні технології в моделюванні» (22-23 березня, 2017 р., м. Миколаїв), XIII Міжнародна конференція «Комп'ютерні науки та інформаційні технології», CSIT 2018, IWIM 2018 (11-14 вересня, 2018 р., м. Львів).

Публікації. За темою дисертації опубліковано 43 наукові праці, серед яких 3 колективні монографії, 1 навчально-методичний посібник з грифом МОН, 20 статей у фахових виданнях України (3 статті у виданнях, які входять до наукометричних баз даних) та 3 свідоцтва про реєстрацію авторського права на

твір, 4 публікації – в іноземних наукових періодичних виданнях, 3 публікації – в іноземних наукових неперіодичних виданнях, 12 матеріалів та тез доповідей на конференціях.

Структура дисертації. Робота складається із вступу, шести розділів, висновків, списку використаних джерел з 244 найменувань та чотирьох додатків. Загальний обсяг роботи становить 335 сторінок, з яких 285 сторінок основного тексту, що дорівнює 12 облікових аркушів, ілюстрованих 62 рисунками та 44 таблицями.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** описано загальну характеристику та обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету та завдання дослідження, описано основні результати, визначено їх наукову новизну та практичне значення.

У першому розділі приведено основні визначення, проведено аналіз існуючих підходів, методів, технічних та програмних засобів для розв’язання поставлених в роботі задач.

Досліджено процес пошуку інформації з боку пошукових систем та користувача. Розглянуто підходи до підвищення ефективності пошуку інформації в мережі Інтернет по ключовим словам. Обґрунтовано доцільність розробки альтернативних моделей ранжування для персоналізації пошукової видачі.

Досліджено актуальні проблеми розвитку глобальної мережі. Розглянуто сучасні програмні засоби для обробки веб-даних та роботи з семантичною розміткою.

Обґрунтовано доцільність застосування комплексного підходу для вирішення проблеми підвищення якості інтелектуального пошуку та обробки даних в мережі Інтернет із застосуванням програмних пошукових агентів, що об’єднує нові методи, моделі та засоби оперування веб-даними на семантичному рівні в режимі реального часу.

У другому розділі описано розроблену систему комплексного оперування веб-даними в мережі Інтернет (КОДІ).

В розділі обґрунтовано переваги застосування комплексного підходу до розв’язання задач дослідження, порівняно з розв’язанням кожної задачі окремо, не розглядаючи їх складовою частиною загальної моделі оперування даними в глобальній мережі.

Система КОДІ є набором окремих модулів, що можуть взаємодіяти між собою. В основі ядра лежить розроблена предметно-орієнтована мова оперування даними в мережі Інтернет (ОДМІ), яка реалізує набір базових функцій обробки та перетворення даних.

Таке архітектурне рішення дозволяє виділити ядро системи в якості рівня абстракції, що з одного боку забезпечує всі модулі системи необхідним набором програмних функцій, а з іншого – приховує деталі їх реалізації. Подібні

архітектурні рішення лежать в основі реалізації моделі вводу-виводу на основі потоків байтів для операційних систем сімейства Unix.

Модулі системи логічно розділені на дві групи:

1. Базові модулі: видобування даних, зберігання даних, обробка даних, відображення даних.

2. Функціональні модулі:

- створення веб-сторінок з інтегрованою семантичною розміткою;
- інтеграція семантичної розмітки в HTML-код вже існуючих веб-сторінок;
- відображення вмісту на основі семантичного профілю веб-ресурсу (шаблони відображення);
- додаткова обробка результатів видачі пошукових систем: метапошук, ранжування результатів пошуку, пошукові агенти.

На рис. 1 представлена загальна схема взаємодії компонентів системи КОДІ.

Базові модулі наслідують необхідні функції ядра системи і додають рівень абстракції між ядром та функціональними модулями. Таке архітектурне рішення реалізує рівневий підхід до проектування систем, де кожен рівень обслуговує свою частину процесу взаємодії. Завдяки такій структурі спільна робота базових та функціональних модулів стає набагато простішою, прозорішою й зрозумілішою.

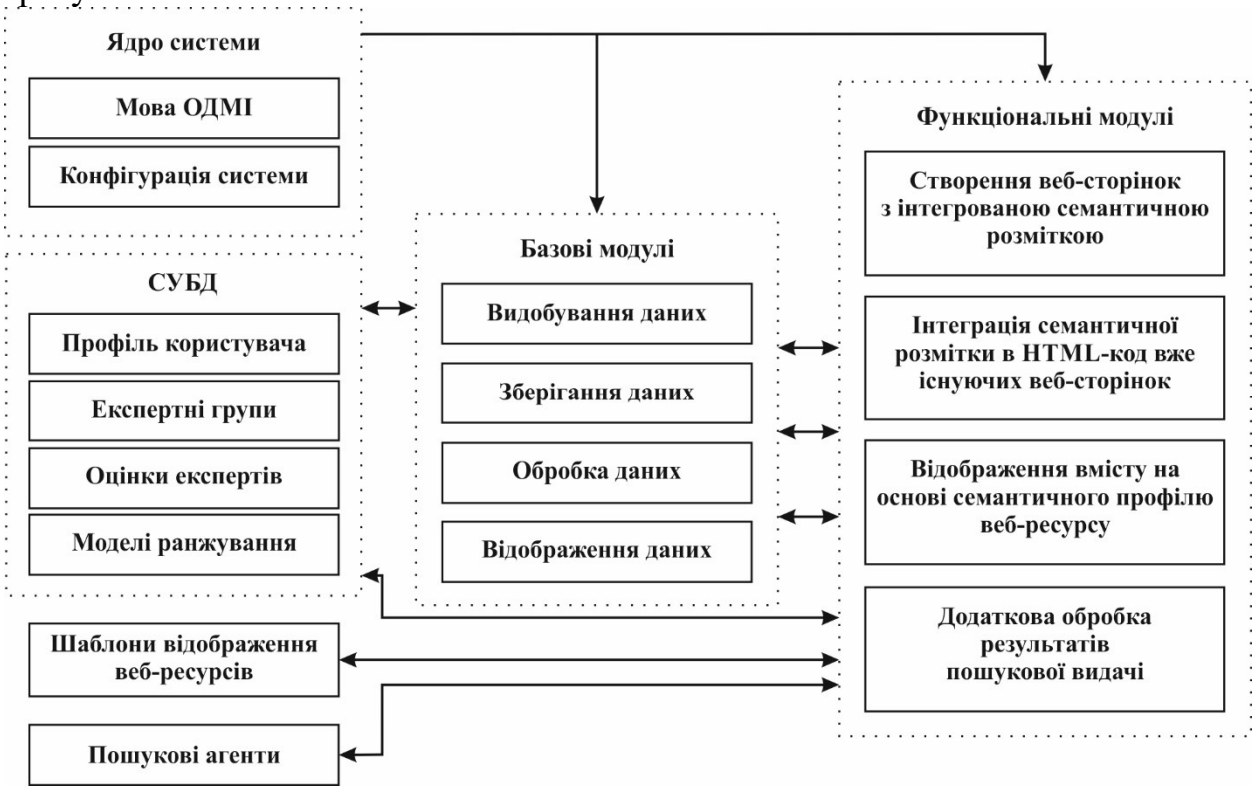


Рис. 1. Схема взаємодії компонентів системи КОДІ

Функціональні модулі реалізують прикладні функції системи у вигляді веб-додатків для забезпечення інформаційних потреб кінцевого користувача. Істотною перевагою використання веб-додатків є коректна робота програми незалежно від операційної системи клієнта.

Функціональні модулі забезпечуються сервісами, які надають базові модулі та ядро системи. Окремі функціональні модулі можуть надавати сервіси іншим. Наприклад, семантичні профілі використовуються при відображенні веб-ресурсів, та результатів пошуку.

Кожен функціональний модуль має власні візуальний та текстовий інтерфейси для взаємодії клієнта з системою.

Домінуючою концепцією у створенні розподілених мережних додатків є архітектура клієнт-сервер, тому саме вона лягла в основу побудови системи КОДІ. Серверна частина містить наступні компоненти:

- базові модулі;
- інтерпретатор предметно-орієнтованої мови;
- база даних;
- система контролю версій;
- обчислювальна частина функціональних модулів.

Клієнтська частина відповідає за взаємодію з користувачем і містить інтерфейси функціональних модулів.

Модуль видобування даних э одним із найважливіших в системі, оскільки без ефективного видобування даних неможлива її подальша робота.

Основою модулю є розроблена предметно-орієнтована мова оперування веб-даними, інтерпретатор якої є частиною ядра системи.

Архітектура модулю проектувалась таким чином, щоб максимально абстрагувати процес пошуку та видобування даних від типу платформи, на якій розроблявся веб-ресурс.

Для пошуку та виділення необхідних фрагментів коду використовуються вбудовані можливості браузера, а саме, консоль розробника, що надає доступ до структури веб-ресурсу. Видобування даних може бути проведено одноразово, або створено шаблон видобування даних із можливістю налаштування розкладу проведення певних операцій.

Модуль збереження даних передбачає використання декількох способів зберігання даних:

- база даних;
- система контролю версій VCS (Version Control System);
- текстовий формат DSV (Delimiter-Separated Values – значення, розділені роздільником).

Модуль обробки даних передбачає три варіанти застосування таких програм:

1. Вбудовані програмні функції системи.
2. Користувацькі програми, додані в систему в якості шаблонів обробки інформації.
3. Зовнішні програмні рішення.

Користувацькі програми для обробки даних можуть бути виконані засобами вбудованої мови ОДМІ, або засобами однієї з мов програмування загального призначення, таких як PHP, або Javascript.

Використання зовнішніх програмних рішень для здійснення операцій над даними, потребує застосування додаткових заходів для забезпечення:

- безпеки і цілісності передачі та отримання даних із зовнішніх програмних продуктів;
- коректності результатів обробки даних.

Ці функції виконує спеціальний контролер обміну даними, через який проходить вся взаємодія модулю обробки даних із зовнішніми програмними продуктами.

Оскільки застосування системи КОДІ передбачає обробку веб-даних, тому всі результати роботи системи представляються кінцевому користувачу у вигляді веб-сторінок. Для відображення даних в системі КОДІ використовується підхід, заснований на шаблонах. При цьому використовуються два види шаблонів для відображення двох груп результатів роботи системи:

1. Шаблони, засновані на семантичному профілі веб-ресурсу. Цей тип шаблонів є основним для системи і застосовується для відображення веб-ресурсів і результатів пошуку.

Кінцева веб-сторінка формується на основі трьох компонентів:

- безпосередньо дані;
- HTML-шаблон, що включає структуру веб-сторінки та стильове оформлення;
- семантичний профіль, згідно якого в HTML-шаблон інтегрується семантична розмітка, необхідна, наприклад, для ефективного та коректного застосування методів інтелектуального пошуку.

2. Шаблони відображення результатів агрегації та обробки даних, таких як таблиці, списки, діаграми, тощо. Описувати такі дані згідно семантичного профілю корпоративних веб-ресурсів важко, а іноді неможливо, тому що вони можуть носити абстрактний характер, не прив’язаний до конкретних товарів або послуг. Наприклад, дослідження динаміки змін погодних умов, формування списків художньої продукції, таких як фільми, книги, аудіо записи, тощо.

Застосування таких шаблонів базується на підході, реалізованому в шаблонах FastTemplate на мові програмування PHP. Будь-яка велика сторінка складається з безлічі фрагментів, найменші з яких - звичайні рядки тексту, які отримують ім'я і значення.

Відображення інформації на основі шаблонів є одним із найбільш вживаних в наш час методів. Під час розробки веб-ресурсу створюється шаблон відображення даних, згідно якого для всіх користувачів мережі даний веб-ресурс відображається однаково. Фактори, що впливають на зовнішній вигляд шаблону:

- побажання замовника;
- особистий досвід розробника;
- тенденції дизайну веб-ресурсів, що домінують в даний момент часу.

Очевидно, що один шаблон не може враховувати особливості сприйняття інформації всіх потенційних користувачів. Це призводить до того, що у певної групи користувачів можуть виникати труднощі із пошуком необхідної інформації.

Основною відмінністю підходу, реалізованого в системі КОДІ є те, що шаблони відображення застосовуються не на стороні розробника веб-ресурсу, а на стороні користувача. Тобто, всі веб-ресурси, створені на основі представленого в роботі семантичного профілю корпоративних веб-ресурсів, можуть бути відображені однаково згідно одного користувацького шаблону.

Схема на рис. 2 демонструє відмінності загальноприйнятого підходу до відображення вмісту веб-ресурсу від запропонованого в роботі.

Умовні позначення на рис. 2: $S_1,\ldots S_n$ – веб-ресурси, $T_1,\ldots T_n$ – шаблони відображення, $U_1,\ldots U_n$ – користувачі.

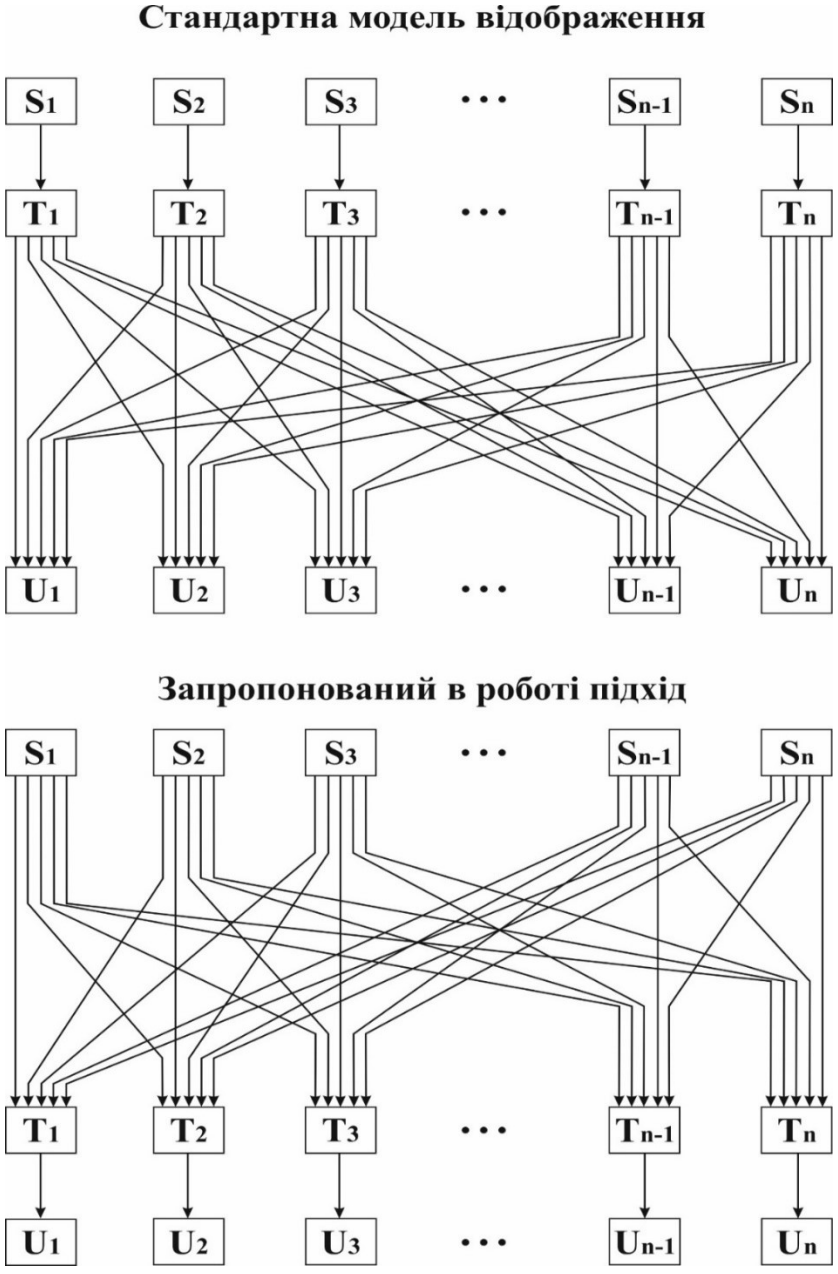


Рис. 2 Схема застосування шаблонів відображення веб-ресурсів.

Відображення всіх веб-ресурсів одного типу згідно одного шаблону з визначеною структурою розташування елементів та їх стильовим оформленням має значні переваги перед існуючим підходом:

1. Користувач не витрачає час на ознайомлення зі структурою нового веб-ресурсу, пошук навігаційного меню та блоку з основним вмістом сайту. Всі типові блоки завжди розташовані на одних і тих же місцях.

2. Користувач має можливість налаштування стильового оформлення веб-ресурсу згідно власних побажань та індивідуальних особливостей – розмір тексту, кольорова гамма, розміри та розташування блоків, тощо.

3. Перші два пункти є очевидними, але потенційні можливості системи набагато ширші. Відсутність жорстко заданих шаблонів з боку розробника веб-ресурсу дає можливість користувачу змінювати структуру веб-ресурсу в цілому. Наприклад, є можливість відключення відображення будь-яких блоків, а також виведення інформації з різних сторінок веб-ресурсу на одному екрані, що значно прискорює процес сприйняття інформації та прийняття рішень.

Даний модуль є базою для створення програмних засобів з вбудованою можливістю інтеграції семантичної розмітки в HTML-код нових веб-сторінок. Запропонований підхід передбачає створення окремих семантичних профілів для найбільш вживаних типів веб-ресурсів: корпоративні, інтернет-магазини, новинні, блоги, інформаційні, форуми, тощо.

В дисертаційній роботі реалізація такого підходу розглядається на прикладі корпоративних веб-ресурсів, як ефективного інструменту підвищення потенціалу розвитку вітчизняного сегменту електронної комерції.

Семантичний профіль не є жорсткою структурою і може бути доповнений новими довільними елементами, або елементами із базового набору словника schema.org.

Архітектура модулю спирається на вдалі рішення системи керування вмістом WordPress, що на сьогоднішній день займає найбільшу долю світового та вітчизняного ринку розробки корпоративних веб-ресурсів. Візуальний інтерфейс забезпечує зручне додавання всіх необхідних даних згідно структури корпоративного веб-ресурсу, що зберігаються в базі даних. При генерації кінцевого HTML-коду веб-сторінок, інформація з бази даних розміщується у відповідних блоках згідно шаблону відображення з додаванням семантичної розмітки згідно семантичного профілю корпоративного веб-ресурсу.

В останнє десятиріччя стрімкого розвитку набули системи розробки веб-ресурсів, в основі яких лежить генерація кінцевого коду веб-сторінки в режимі реального часу в момент надходження HTTP-запиту від користувача на сервер. Веб-сторінки в готовому вигляді не існують до моменту запиту. Відповідно, немає можливості додати семантичну розмітку в код веб-сторінки в ручному режимі. Генерація HTML-коду сторінки відбувається на основі шаблону відображення, який описує HTML-структуру майбутньої сторінки, а інформаційне наповнення зберігається в базі даних. Додавання функції інтеграції семантичної розмітки в кінцевий код веб-сторінок в подібних системах неможливе без значного втручання в початковий код самої системи. Для деяких з них існують часткові рішення цієї задачі, реалізовані у вигляді плагінів та модулів. Вони використовуються в основному для додавання розширених сніппетів і не дають можливості повноцінної семантичної розмітки веб-ресурсу.

Враховуючи все вищезазначене, інтеграція семантичної розмітки в HTML-код вже існуючих веб-сторінок є актуальною і складною задачею.

Розв'язання цієї задачі зовнішніми методами без втручання в роботу системи генерації веб-сторінок потребує наявності наступних рішень:

- ефективні методи видобування даних;
- семантичний профіль веб-ресурсу;
- створення та зберігання шаблону семантичної розмітки;
- відображення веб-сторінки з накладанням шаблону семантичної розмітки.

В роботі значна увага приділяється саме цій задачі, оскільки кількість вже існуючих веб-сторінок обчислюється сотнями мільйонів і переважна більшість з них не має семантичної розмітки. Очевидно, що практичне втілення ідей концепції семантичної павутини в існуюче веб-середовище неможливе без створення ефективних та простих у використанні методів інтеграції семантичної розмітки в HTML-код вже функціонуючих веб-ресурсів.

Видобування даних реалізовано у вигляді відповідного модулю системи КОДІ. Розроблений семантичний профіль корпоративних веб-ресурсів цілком забезпечує необхідний набір класів для опису структур. Шаблон семантичної розмітки являє собою метаопис існуючого веб-ресурсу. Він містить набір визначених користувачем класів опису семантичної структури веб-ресурсу з прив'язкою до інформаційного наповнення і зберігається в базі даних. За наявності HTTP-запиту до серверу відбувається генерація HTML-коду і шаблон семантичної розмітки накладається на існуючий шаблон відображення веб-сторінки шляхом включення додаткових атрибутів тегів в HTML-блоки, що безпосередньо містять інформаційне наповнення. Такий формат розміщення семантичної розмітки має назву – мікродані. Оскільки прив'язка семантичних атрибутів застосовується до HTML-тегів шаблону, а не до самих даних, то оновлення інформації на сайті не потребує повторного оновлення семантичної розмітки. Це дуже важливо, оскільки більшість сучасних сервісів для додавання семантичної розмітки до вже існуючих сайтів використовують формат JSON-LD. Його особливість полягає генерації кінцевого коду на мові Javascript у вигляді «фрагмент даних»:«властивість», який вставляється за допомогою тегу <script> на початку HTML-документу. Такий код є статичним і потребує регулярної повторної розмітки фрагментів даних, що були оновлені, в ручному режимі.

Для відображення результатів роботи в системі КОДІ використовується механізм шаблонів. Даний модуль реалізує новий підхід до відображення вмісту веб-ресурсів, в якому шаблони відображення застосовуються на стороні клієнта для всіх веб-ресурсів заданого типу. Такий підхід є альтернативою до традиційного, в якому шаблони відображення жорстко прив'язані до кожного веб-ресурсу.

Необхідною умовою ефективного застосування такого підходу є наявність заздалегідь спроектованої структури та семантичного профілю певного виду веб-ресурсів.

Шаблон включає в себе:

- структурну схему розміщення HTML-блоків веб-сторінки;
- візуальне стильове оформлення зазначених блоків;
- семантичний вміст HTML-блоків.

Окрему увагу слід приділити поняттю семантичного вмісту HTML-блоку. Воно визначає, які саме елементи семантичного профілю будуть відображені в конкретному HTML-блоці. Наприклад, стандартний підхід передбачає логічний поділ веб-ресурсу на сукупність веб-сторінок, кожна з яких має свою унікальну URL-адресу в рамках доменного імені (http://site.ua/pageA ... http://site.ua/pageZ). Звернення до конкретної URL відображає на екрані вміст однієї сторінки. Запропонований в роботі підхід передбачає два варіанти унікальної ідентифікації інформаційного вмісту:

1. Загальноприйнятий, в я кому інформація прив'язана до конкретної сторінки веб-ресурсу. В якості ідентифікатора виступає URL-адреса сторінки.

2. Ідентифікація інформації на основі семантичної розмітки, яка може бути використана без прив'язки до URL-адреси батьківської сторінки. Такий підхід дає можливість, виводити повний або частковий вміст декількох, або всіх сторінок веб-ресурсу на одному екрані.

Користувацькі шаблони можуть бути ефективно застосований не тільки для відображення вмісту веб-ресурсу, а й для відображення результатів пошукової видачі.

Оскільки кожна одиниця інформації має унікальний семантичний ідентифікатор, це дає змогу сформувати шаблон відображення результатів пошукової видачі в довільному вигляді. Зазначений підхід використовується в якості важливої складової частини модулю додаткової обробки результатів видачі пошукових систем.

На сьогоднішній день безперечним світовим лідером в галузі пошуку інформації в мережі Інтернет є пошукова система Google. Тому результати пошукової видачі саме цієї системи взяті за основу подальших досліджень. Виходячи з цього, в роботі основна увага приділяється додатковій обробці результатів пошукової видачі за трьома напрямками:

- розробка альтернативних методів ранжування результатів пошуку;
- створення нових підходів до відображення результатів пошуку;
- удосконалення методів інтелектуального пошуку на базі пошукових агентів.

В роботі представлена альтернативна модель ранжування результатів пошуку на основі оцінок користувачів. Вона була побудована, як один із варіантів розв'язання задачі персоналізації результатів пошуку. Розв'язанню цієї задачі присвячено багато вітчизняних та закордонних досліджень. Пошукові системи постійно розробляють нові методи для підвищення персоналізації результатів пошуку. Але суть моделі ранжування з моменту створення Google принципово не змінилися. Вона все також в більшій мірі спирається на об'єктивні фактори значущості веб-ресурсу, такі як кількість та якість зовнішніх посилань на сторінку, щільність ключових слів на сторінці, тощо.

У представленому в роботі підході до підвищення персоналізації результатів пошуку для кожного користувача формується унікальна експертна група на основі міри узгодженості думок поточного користувача з потенційними експертами. Нові ранги веб-ресурсів обчислюються на основі оцінок експертів з урахуванням ваги кожного експерта відносно поточного користувача. Згідно цієї моделі порядок ранжування результатів пошуку для кожного користувача буде індивідуальним. Персоналізація результатів пошуку важлива для розв'язання наступних проблем сучасного пошуку:

- боротьба з пошуковим спамом;
- підвищення якості пошукової видачі;
- скорочення часу на пошук необхідної інформації.

Удосконалення методів інтелектуального пошуку засновано на застосуванні семантичного профілю в якості ідентифікатора інформації для пошукових агентів. Ефективна робота пошукових агентів стає можливою завдяки комплексному застосуванню низки представлених в роботі підходів та методів:

- семантичний профіль веб-ресурсу;
- метод видобування даних на основі мови ОДМІ;
- зберігання даних із використанням систем контролю версій, що дає широкі можливості для агрегації та подальшого аналізу даних;
- користувацькі шаблони відображення результатів пошуку.

Модуль передбачає програмування та застосування пошукових агентів безпосередньо в процесі роботи з пошуковою видачою.

Для підвищення ефективності запропонованих рішень, всі вони були об'єднані в одну систему КОДІ, що надає веб-розробникам, власникам веб-ресурсів та користувачам глобальної мережі широкі можливості оперування веб-даними для реалізації власних інформаційних потреб.

Система розроблена в рамках концепції семантичної павутини і орієнтована на вдосконалення методів пошуку інформації, створення веб-ресурсів та програмних пошукових агентів.

Третій розділ присвячено розробці семантичного профілю корпоративних веб-ресурсів та створенню на його базі методу відображення інформаційного вмісту на основі користувацьких шаблонів.

В результаті дослідження інформаційного наповнення та елементів навігації однієї тисячі корпоративних веб-ресурсів, було побудовано їх загальну структуру. Аналіз відбувався в два етапи:

1. Видобування та дослідження елементів навігації верхнього рівня.
2. Дослідження інформаційного вмісту для складання внутрішньої структури елементів верхнього рівня.

Видобування елементів навігації здійснювалось на основі побудованого в роботі алгоритму, що враховував їх характерні особливості.

З кожного веб-ресурсу видобувалися елементи навігаційного меню у вигляді «Назва елементу» → «Посилання» і зберігалися в базу даних. В таблиці 1 представлено перелік структурних елементів впорядкованих за зменшенням

частоти їх входжень до веб-ресурсів. Всього збережено 5371 унікальних елементів навігації.

Елементи з числом входжень менше 50 були проігноровані, як неінформативні.

Наступним кроком в ручному режимі було переглянуто 100 обраних довільним чином веб-ресурсів з попередньої вибірки для дослідження інформаційного вмісту базових структурних елементів.

Загальна структура корпоративних веб-ресурсів була побудована, спираючись на результати експерименту, а також на результати досліджень ведучих спеціалістів в галузі веб-дизайну, та зручності використання веб-ресурсів - usability.

Таблиця 1

Перелік структурних елементів корпоративних веб-ресурсів

№ з/п	Назва елементу	Кількість входжень	№ з/п	Назва елементу	Кількість входжень
1	Контакти	891	24	Нагороди	84
2	Про компанію	637	25	Партнери	81
3	Послуги	611	26	Клієнти про нас	78
4	Новини	363	27	FAQ	75
5	Продукція	232	28	Проекти	74
6	Вакансії	217	29	Дилерам	72
7	Про нас	193	30	Документи	72
8	Співпраця	185	31	Наші об'єкти	71
9	Сертифікати	177	32	Акції	68
10	Інформація	169	33	Умови співпраці	65
11	Сервіс	156	34	Для дилерів	63
12	Відгуки	146	35	Наші контакти	62
13	Головна	142	36	Події	60
14	Кар'єра	138	37	Відзнаки	59
15	Статті	125	38	Для партнерів	57
16	Пресс-центр	124	39	Клієнти	56
17	Ліцензії	102	40	Наші проекти	56
18	Наші послуги	98	41	Наша компанія	54
19	Наші роботи	97	42	Наші партнери	53
20	Наша продукція	93	43	Виробництво	52
21	Об'єкти	93	44	Галерея робіт	51
22	Наші клієнти	87	45	Наші новини	50
23	Магазин	86			

Загальна структура корпоративних веб-ресурсів стала базою для побудови семантичного профілю.

В ході аналізу існуючих стандартів семантичної розмітки в якості базової онтології був обраний стандарт schema.org. Він включає найбільшу кількість сутностей, активно розвивається і має великий потенціал для інтеграції нових

спеціалізованих семантичних профілів. Побудова семантичного профілю корпоративних веб-ресурсів складалася з виділення переліку необхідних класів для опису структури зі словника shema.org та подальшого додавання нових семантичних класів, відсутніх в базовій онтології.

На рис. 3 представлена UML-діаграма класів розробленої семантичної моделі корпоративного веб-ресурсу.

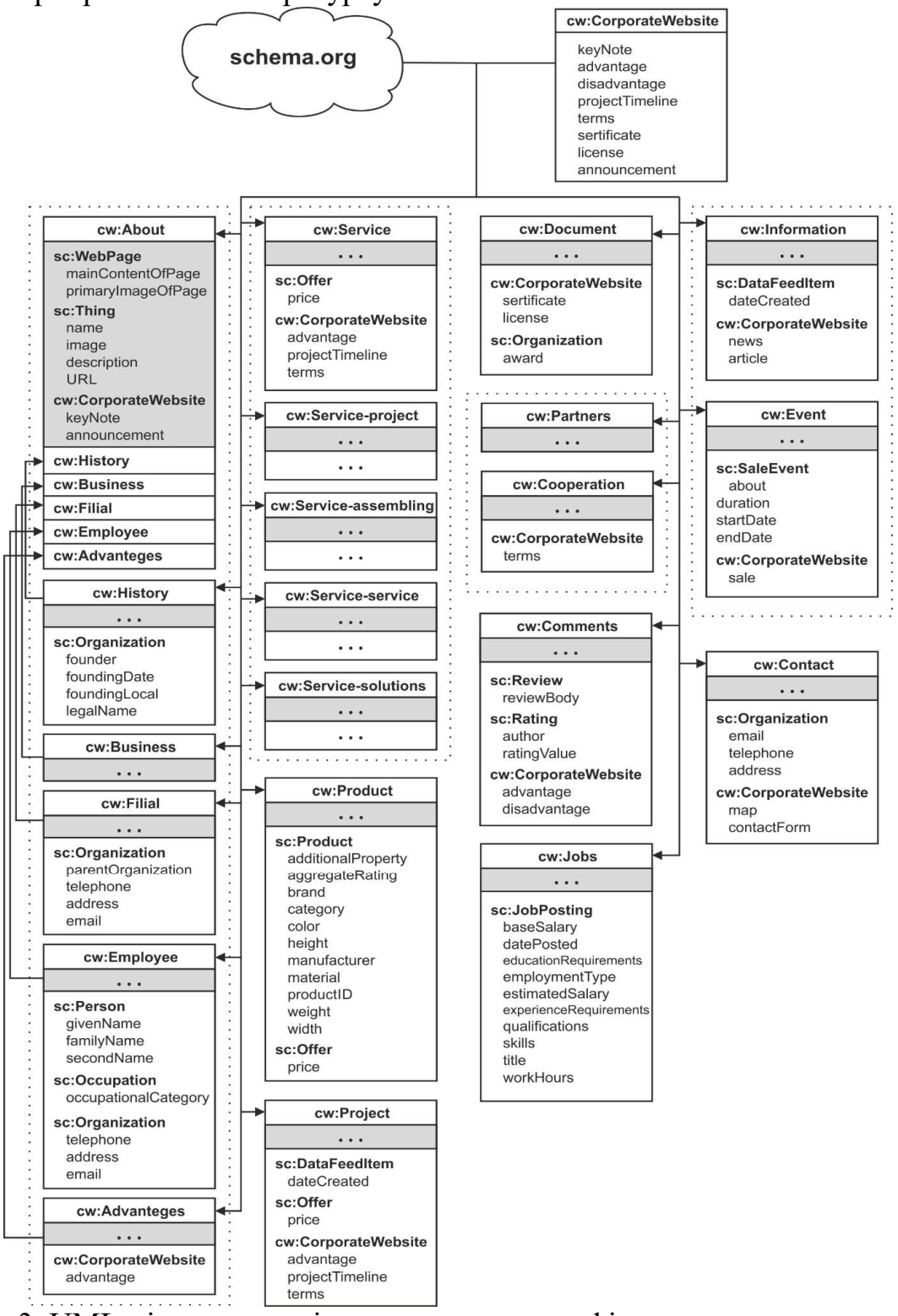


Рис. 3. UML-діаграма класів семантичного профілю корпоративного веб-ресурсу.

Властивості, загальні для кожного класу помічені сірим фоном і для зручності сприйняття у всіх класах окрім першого, замінені на «…».

Класи відповідають структурним елементам першого рівня. Кожен клас наслідує з онтології schema.org та базового класу cw:corparateWebsite, дві групи властивостей:

1. Загальна для всіх класів.

sc:Thing:

- name – назва;
- image – зображення;
- description – опис;
- URL – унікальна адреса вебсторінки.

sc:WebPage:

- mainContentOfPage – основний вміст сторінки;
- primaryImageOfPage – основне зображення на сторінці.

cw:CorporateWebsite:

- keyNote – найважливіша частина вмісту сторінки, ключові слова, основна думка, що може бути використана для підвищення якості пошуку інформації;
- announcement – стислий анонс представленої на сторінці інформації, як правило використовується на сторінці з переліком новин, статей, вакансій, відгуків, тощо.

2. Специфічна для конкретного класу.

Кожен клас має свій специфічний набір властивостей, що описує інформаційний вміст сторінок даного типу.

Новий клас cw:CorporateWebsite має наступні властивості:

- keyNote;
- advantage – перевага, деяка властивість, що вигідно відрізняє об’єкт або послугу від інших;
- disadvantage – недолік;
- projectTimeline – строк виконання проекту;
- terms – умови;
- sertificate – сертифікат;
- license – ліцензія;
- announcement.

Побудовані структурна схема та семантичний профіль корпоративних веб-ресурсів є базою для розробки наступних елементів системи КОДІ:

1. Спеціалізована система керування вмістом корпоративних веб-ресурсів з вбудованими функціями автоматичної інтеграції семантичної розмітки в HTML-код веб-сторінок.

2. Модуль відображення вмісту веб-ресурсів на основі користувацьких шаблонів.

3. Метапошукова система, заснована на обробці пошукової видачі популярних пошукових систем із застосуванням пошукових агентів та можливістю відображення результатів пошуку на основі шаблонів користувачів.

4. Персоналізована веб-сторінка користувача, на якій відображається актуальна для користувача інформація, яка автоматично видобувається і агрегується з різних джерел за допомогою пошукових агентів.

Наступним кроком розв'язку задач дослідження є розробка методів інтеграції семантичної розмітки в HTML код вже існуючих веб-сторінок, а також розробка системи створення веб-ресурсів з інтегрованою семантичною розміткою.

Розробка семантичного профілю корпоративного веб-ресурсу відкриває нові можливості для розвитку галузі вітчизняної електронної комерції.

Практичне втілення концепції семантичної павутини потребує розв'язання двох актуальних задач:

1. Інтеграція семантичної розмітки до вже існуючих веб-ресурсів.

2. Розробка нових веб-ресурсів з інтегрованою семантичною розміткою.

Розв'язання цих задач потребує розробки стратегії розмітки веб-ресурсу, що включає в себе:

- виділення елементів, які необхідно розмітити;
- вибір стандарту, згідно якого буде здійснюватися розмітка;
- вибір автоматичного, або ручного підходу до інтеграції розмітки;
- вибір інструментів генерації коду мікророзмітки;
- вибір способу та інструментів інтеграції мікророзмітки в HTML-код веб-ресурсу.

За способами, додавання семантичної розмітки поділяють на безпосередню інтеграцію розмітки в якості атрибутів HTML-тегів, та окреме додавання розмітки на початку HTML-документу в якості фрагменту javascript-коду.

Використання підходу, що передбачає додавання семантичної розмітки у вигляді окремого фрагменту коду, є більш простим та зручним для одноразового використання. Але цей підхід має значний недолік – відсутність можливості автоматичної актуалізації семантичної розмітки при оновленні інформації на сайті.

Інтеграція семантичної розмітки в якості атрибутів HTML-тегів веб-сторінки є набагато більш трудомістким процесом, що вимагає професійних знань в галузі розробки веб-додатків. Основна його перевага полягає у відсутності необхідності оновлювати семантичну розмітку щоразу після внесення змін на сайті. Але, не дивлячись на очевидну перевагу, цей підхід майже не застосовується через складність реалізації та відсутність програмних засобів для автоматизації процесу.

Дослідження методів розробки веб-ресурсів показало значні переваги використання системи керування вмістом, CMS (Content Management System). Для найбільш ефективної реалізації задач дослідження, було прийнято рішення розробити нову систему керування вмістом на основі розробленої структури корпоративних веб-ресурсів з інтегрованою семантичною розміткою.

В запропонованій системі керування вмістом представлені всі типи веб-сторінок у відповідності до розробленої структури. Кожний тип сторінки

включає власний набір полів. Кожне поле для вводу інформації зв'язано з певним класом семантичної розмітки, що дозволяє генерувати код з вбудованою семантичною розміткою.

Семантичний профіль веб-ресурсу будується у вигляді шаблону відображення основних блоків веб-сторінки з додаванням семантичної розмітки.

Метод застосування користувацьких шаблонів передбачає одночасне відображення HTML-блоків з різних сторінок веб-ресурсу на одному екрані. По аналогії з користувацькими шаблонами відображення веб-ресурсів, в роботі реалізовано шаблони відображення результатів пошуку.

Сучасні пошукові системи, зокрема Google, використовують один шаблон для відображення всіх результатів пошукової видачі (рис. 4).

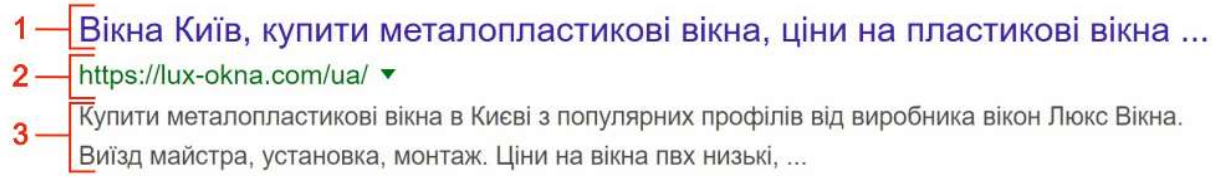


Рис. 4. Результат пошукової видачі Google

Шаблон складається з трьох блоків:

1. Заголовок, містить текст заголовку із тегу <title>…</title>.

2. Посилання на веб-сторінку, що відповідає пошуковому запиту.

3. Текстовий блок з описом вмісту веб-сторінки. Блок містить текст із метатегу <meta name=«description» content=«…»>.

Призначення тегів <title> та <meta name=«description»>, закладене консорціумом W3C при їх створенні - надання точної та достовірної інформації про вміст веб-сторінки. Відповідно, пошукові системи при ранжуванні результатів спираються на наявність в цих тегах ключових слів, заданих в пошуковому запиті, як на основний, найбільш важливий фактор. В результаті чого ці теги перетворились на один з основних інструментів штучної маніпуляції результатами пошукової видачі. Замість достовірної та точної інформації про вміст веб-сторінки, в них розміщується інформація, що відповідає найбільш популярним серед користувачів мережі пошуковим запитам.

Інформація, що описує результати пошуку є недостатньою для прийняття рішення про те, чи є представлена на веб-ресурсі інформація корисною для відвідувача. Причому пошукова система не надає ніяких інструментів для редагування шаблону відображення результатів.

Застосування методу відображення результатів пошуку на основі користувацьких шаблонів має наступні переваги:

- частково розв'язує проблему штучного просування;
- скорочує час на пошук необхідної інформації.

У четвертому розділі представлено розроблену мову оперування даними в мережі Інтернет.

Метою розробки предметно-орієнтованої мови було забезпечення представленої в роботі технології зручними та інтуїтивно зрозумілими програмними засобами обробки веб-даних, що будуть доступні не тільки

досвідченим розробникам, а й власникам веб-ресурсів, які не мають практичних навичок в галузі програмування.

Обробка веб-даних пов’язана з виконанням великої кількості однотипних операцій (видобування, збереження, обробка, відображення даних) з різними наборами даних, представлених на веб-ресурсах з різною структурою. Для реалізації програмних інструментів оперування веб-даними на мовах програмування високого рівня загального призначення, необхідно витрачати велику кількість часу на адаптацію програмного коду з урахуванням структури конкретного веб-ресурсу. До того ж, такий підхід не доступний недосвідченим користувачам.

Резлізація граматики та лексики мови ОДМІ здійснювалась на базі проекту Marpa - реалізації алгоритму Ерлі з урахуванням оптимізацій Лео, рішенням помилки з правилами нульової довжини і власними оптимізаціями для поліпшення локалізації помилок.

Граматика

Code ::= Declaration+
Declaration ::= Fetch | Assignment | Navigation | Auth | Configuration
Fetch ::= FETCH URL
Assignment ::= NAME '=' SELECTOR
Navigation ::= NAV SELECTOR
Navigation ::= NAV SELECTOR Navigation_attribute
Navigation_attribute ::= NAV_ATTR_single | NAV_ATTR_form | NAV_ATTR_block
Auth ::= AUTH Auth_form
Auth ::= AUTH Auth_selectors
Auth_selectors ::=Auth_selector+
Auth_selector ::= SELECTOR '=' ATTR_VALUE
Auth_form ::= LOGIN '=' ATTR_VALUE PASSWORD '=' ATTR_VALUE
Auth_form ::= PASSWORD '=' ATTR_VALUE LOGIN '=' ATTR_VALUE
Configuration ::= CONF Configuration_pairs
Configuration_pairs ::= Configuration_pair+
Configuration_pair ::= CONF_ATTR '=' ATTR_VALUE

Лексика

CONF_ATTR ~ '_' CONF_ATTR_NAME '_'
CONF_ATTR_NAME ~ [\w]+
ATTR_VALUE ~ ["] dquoted ["]
dquoted ~ [^"]+
FETCH ~ 'fetch':i
FETCH ~ 'get':i
FETCH ~ 'взяти':i
NAV ~ 'nav':i
AUTH ~ 'auth':i
AUTH ~ 'вхід':i

```
LOGIN ~ 'user':i
LOGIN ~ 'користувач':i
PASSWORD ~ 'password':i
PASSWORD ~ 'passwd':i
PASSWORD ~ 'pass':i
PASSWORD ~ 'пароль':i
CONF ~ 'conf':i
CONF ~ 'конф':i
NAV_ATTR_single ~ 'single':i
NAV_ATTR_single ~ 'один':i
NAV_ATTR_form ~ 'form':i
NAV_ATTR_form ~ 'форма':i
NAV_ATTR_block ~ 'block':i
NAV_ATTR_block ~ 'блок':i
URL ~ url_prefix url_body
url_prefix ~ url_prefix_http | url_prefix_https
url_prefix_http ~ 'http://':i
url_prefix_https ~ 'https://':i
url_body ~ [\w!$&()*+,;=:/?#\[\]@%]+
NAME ~ id | array
id ~ [\w\._]+
array ~ '@' id
SELECTOR ~ ['] squoted [']
squoted ~ [^']+
:discard ~ whitespace
whitespace ~ [\s]+
```

Важливо відзначити, що всі ключові слова мови ОДМІ можуть бути переведені на іншу мову, а також те, що вони не є зарезервованими на відміну від багатьох інших мов і, при необхідності, будуть використані як імена змінних. У наведеному фрагменті для всіх слів є як англійські так і українські варіанти, а для деяких дано кілька найбільш поширених варіантів на одній мові. Подібний підхід значно зменшує кількість помилок при написанні програмного коду.

Типи оголошень і їх семантика перераховані нижче:

Присвоєння (Assignment). Кожне оголошення або починає нове завдання, або додає дані для вже наявного. Крім цього, кожне присвоєння може додати новий елемент в структуру Крім вказівки потрібного блоку може виконуватися і додаткова обробка отриманих даних. У поточній реалізації це тільки видобування тексту або залишення у вигляді HTML. Однак можливе додавання інших обробників.

Завантаження (Fetch) URL. Кожен URL є ключем в таблиці завдань. Якщо ключ відсутній, то створюється нове завдання. Якщо присутній, то вибирає його в якості поточного.

Навігація (Navigation). Вказує способи отримання нових сторінок у разі, якщо дані розбиті по ним, а не знаходяться в межах однієї. Не починає нове

завдання, але додає неявний цикл для поточного. Не може бути вкладеним, завжди працює починаючи від URL завдання. Одне завдання може включати в себе кілька навігацій.

Аутентифікація (Auth). Може бути тільки одна. Вказує спосіб аутентифікації для веб-ресурсів, які її потребують для видачі даних. Виконується до будь-яких завдань.

Конфігурація (Configuration). Може бути тільки одна. Задає значення ряду параметрів для документа в цілому. В основному для системи зберігання.

Для ілюстрації переваг предметно-орієнтованих мов, було програмно реалізовано розв'язання задачі видобування даних з HTML-сторінки веб-ресурсу bezpeka.top про наявні системи сигналізації та збереження їх в базу даних.

Нижче представлено приклад розв'язання поставленої задачі засобами мови програмування високого рівня загального призначення Perl із застосуванням спеціалізованих бібліотек для обробки веб-даних.

```
#!/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="сигнализації";
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 реалізації, отриманий програмний код є важким для розуміння недосвідчених користувачів. Для будь-яких модифікацій спеціалістам з предметної галузі доведеться вивчити велику кількість документації стосовно використаних модулів і мати базові навички роботи з мовою Perl.

Крім того дана програма має наступні недоліки:

- наявність лише базової обробки помилок;
- нестійка до модифікацій API використаних модулів;
- 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.

Для порівняння, та ж сама задача була розв’язана засобами розробленої предметно-орієнтованої мови.

```
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
```

Така програмна реалізація є набагато більш виразною. Більшу частину програмного коду становить інформація про структуру представлення даних на веб-ресурсі, яку користувач легко може отримати використовуючи вбудовані

функції сучасних браузерів. Такий код легко писати, читати і модифікувати недосвідченим користувачам. Його не потрібно модифікувати у разі використання різних видів сховищ даних для збереженні інформації, або використання іншого алгоритму видобування та парсингу HTML-коду. Завдання обробки помилок, налагодження і тестування бере на себе система виконання.

На рис. 5 зображено структурну схему взаємодії компонентів мови ОДМІ.

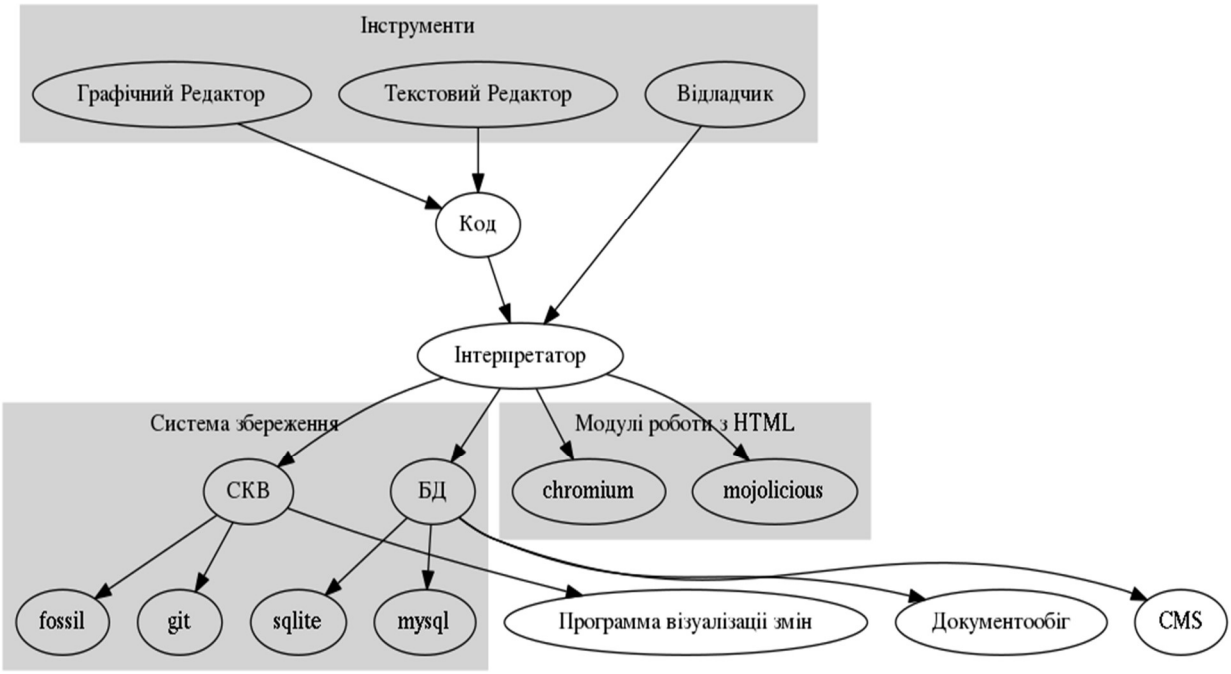


Рис. 5. Схема взаємодії компонентів предметно-орієнтованої мови оперування веб-даними.

П’ятий розділ присвячено розробці персоналізованої моделі ранжування результатів пошуку на основі оцінок експертів, думка яких є авторитетною для користувача.

Для визначеності в означеннях, в рамках дослідження пропонуються ввести такі терміни:

Поточний користувач – користувач, відносно якого формується група експертів і для якого відбувається ранжування результатів пошуку.

Вага експерта – міра узгодженості думок експерта та поточного користувача, розрахована на основі тісноти зв‘язку їх оцінок для деякої множини веб-ресурсів. Вагу експерта U_i відносно поточного користувача U_0 доцільно вважати метрикою $d(U_0, U_i)$ на метричному просторі (X, d), де $X[U_0,\ldots,U_n]$ є множина всіх користувачів системи. Функція d задовольняє умовам аксіом тотожності, симетрії та трикутника, однак вона не може бути визначена для кожної пари елементів з множини X, так як не всі користувачі мають спільні оцінки. Отже в контексті даної задачі некоректно використовувати термін метричного простору, тому в подальшому функцію $d(U_0, U_i)$ визначення ваги експерта будемо називати аналогом метрики.

Метод середніх різниць оцінок (СРО) – запропонований в роботі метод розрахунку ваги експертів.

Потенційний експерт першого рівня U_i – користувач, що має спільні оцінки з поточним користувачем і для якого ще не розрахована міра узгодженості думок.

Потенційний експерт другого рівня $\hat{U}_j$ – користувач, що не має спільних оцінок з поточним користувачем, але має спільні оцінки з експертом першого рівня.

Потенційний експерт третього рівня $\widetilde{U}_k$ – користувач, що не має спільних оцінок з поточним користувачем та експертами першого рівня.

Для пошукового запиту поточного користувача, ранги веб-ресурсів з пошукової видачі, обчислюються як середнє гармонічне зважене всіх оцінок членів експертної групи. Експертні групи для кожного користувача мережі є унікальними і формуються у фоновому режимі роботи системи КОДІ за трьома методами в залежності від наявності спільних на деякій множині веб-ресурсів оцінок між поточним користувачем та потенційним експертом. Експерти поділяються на три рівня, відповідно до методу розрахунку їх ваги. Поділ на рівні має лише логічний характер. Вага експертів всіх рівнів рівнозначна і враховується при ранжуванні результатів пошуку без додаткових коефіцієнтів.

Вага експертів першого рівня розраховується за методом середніх різниць оцінок, який включає:

- розрахунок середнього значення модулів різниць оцінок для кожної пари «поточний користувач – потенційний експерт»

$$d(U_0, U_i) = \sum_{j=1}^{m} \frac{\left| U_{0j} - U_{ij} \right|}{m}, \qquad (1)$$

де m – кількість спільних оцінок, що мають поточний користувач з i-тим потенціальним експертом з множини X, j – порядковий номер оцінки

- застосування нормуючої функції для приведення даних в шкалу від 0 до 0,99.

$$W(d) = 1 - \left(1{,}1 \cdot \frac{d(U_0, U_i)}{10} \right) \qquad (2)$$

- відбір користувачів, що мають високу силу зв’язку з поточним користувачем за школою Чеддока (значення >0,7).

Результати досліджень показали, що для ефективного розв’язання задачі, доцільно використовувати саме метод середніх різниць оцінок (СРО), який дає значно коректніші результати ніж метод конкордації Кендала і подібні йому. Сетод СРО може приймати на вхід значення оцінок в порядковій шкалі, в той час як інші методи приймають в якості вхідних даних список рангів.

Вагу експертів другого рівня відносно поточного користувача, доцільно розраховувати як добуток ваги експерта першого рівня відносно потенційного експерта другого рівня, що мають спільні оцінки, та ваги експерта першого рівня відносно поточного користувача:

$$W(d(U_0,\hat{U}_j)) = W(d(U_i,\hat{U}_j)) \cdot W(d(U_0,U_i)) \quad (3)$$

До експертної групи відбираються потенційні експерти другого рівня, значення ваги яких відносно поточного користувача $W(d(U_0,\hat{U}_j)) > 0{,}7$ за шкалою Чеддока.

Вага експертів третього рівня розраховується у випадку відсутності спільних оцінок, на основі соціального профілю користувача.

Для побудови моделі розрахунку ваги експертів на основі соціального профілю, було обрано ряд суб'єктивних ознак x_m, які можуть прямо або побічно впливати на оцінку відвідувача:

x_1 – активність експерта, що виражається в загальній кількості оцінок;

x_2 – різниця у віці, що на відміну від абсолютного значення повних років, є більш суб'єктивним показником відносно користувача. Співпадає з користувачем – 0,9, кожен рік різниці – (-0,01);

x_3 – дохід. Різниця в тисячах гривень;

x_4 – відповідь на питання 3+3*3=? Вірна відповідь – 0,9, невірна – 0,1;

x_5 – стать. Співпадає з користувачем – 0,9, не співпадає – 0,1;

x_6 – наявність домашніх тварин. Співпадає – 0,9, не співпадає – 0,1;

x_7 – освіта. Різниця між освітою користувача та експерта, виражена в числовому значенні. Шкала оцінювання освіти від 1 до 6, де 1 - середня освіта, і 6 - вчений ступінь;

x_8 – фах освіти, гуманітарна, технічна. Співпадає – 0,9, не співпадає – 0,1;

x_9 – сімейний стан. Співпадає з користувачем – 0,9, не співпадає – 0,1;

x_{10} – наявність автотранспортного засобу. Співпадає з користувачем – 0,9, не співпадає – 0,1;

x_{11} – заняття спортом. Співпадає з користувачем – 0,9, не співпадає – 0,1;

В таб. 2 представлені зведені результати моделювання за модифікаціями алгоритмів методу групового урахування аргументів та з використанням нейромережі.

Порівняльний аналіз методів індуктивного моделювання та нейромережевого методу показав, що нейромережа дає найбільш точні результати. Однак задача полягала в побудові моделі залежності міри узгодженості думок від соціально-особистісних факторів користувачів, а нейромережевий метод не дає можливості отримання такої моделі. Краща модель була знайдена за допомогою узагальненого ітераційного алгоритму, який містить в собі всі попередні ітераційні структури.

Одним з найпоширеніших методів ранжування на основі оцінок є метод середніх балів. Традиційно, застосування методу середніх балів передбачає використання середнього арифметичного для розрахунку значень.

Однак вважається, що такий спосіб є некоректним, оскільки бали зазвичай виміряні в порядковій шкалі, і обґрунтованим є використання медіан рангів для визначення середніх балів.

Для перевірки ефективності та коректності застосування представлених вище методів для розв'язання задачі ранжування веб-ресурсів на основі оцінок

користувачів було проведено 4 експерименти. Було сформовано 4 вибірки даних, по одній на кожен експеримент. В кожній вибірці 20 користувачів оцінювали 20 веб-ресурсів по якості представленої інформації та зручності використання.

Таблиця 2

Зведені результати моделювання ваги експертів.

Алгоритм	R^2, %	Наявність істинних одночленів											**Число зайвих одночленів**
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	
Combi	59,42	+		+	+	+		+			+	+	4
Ітераційні алгоритми з лінійним частинним описом моделі													
БІА	42,2	+	+	+		+	+	+	+	+			3
РІА	17,7			+									10
КІА	55,14	+	+	+		+	+	+	+	+	+		2
Ітераційні алгоритми з квадратичним частинним описом моделі													
БІА	56,25			+		+	+	+	+	+			5
РІА	54,25					+		+		+			8
КІА	77,17			+		+		+					8
БІКА	71,8	+		+		+	+	+	+		+		4
РІКА	54,25					+		+		+			8
УІА	80,27	+	+			+		+	+				6
NN	81,96	NA											NA

В роботі для обчислення середніх балів використовувались наступні способи: середнє арифметичне (СА), середнє арифметичне зважене (САЗ), середнє гармонічне (СГ), середнє гармонічне зважене (СГЗ).

Зважені модифікації середнього арифметичного та середнього гармонічного дозволяють розрахувати нові ранги веб-ресурсів з урахуванням ваги експертів.

По результатах експериментів можна зробити наступні висновки:

1. Метод медіан рангів не може використовуватись для розв’язання задачі ранжування веб-ресурсів на основі оцінок користувачів тому що:

- може приймати в якості вхідних даних лише ранги досліджуваних об’єктів від 1 до n, де n дорівнює їх кількості, в той час як на вхід подаються оцінки за шкалою від 1 до 10;
- не враховує вагу експертів.

2. Результати, додаткових досліджень доводять ефективність урахування ваги експертів при розрахунках рангів веб-ресурсів та обґрунтовують необхідність відсіву користувачів з низькими показниками міри узгодженості думок відносно поточного користувача.

Для оцінки ефективності зазначених вище способів розрахунку середніх балів, використовується середнє відхилення від еталонного ранжування. Воно розраховується як середнє арифметичне суми різниць позицій веб-ресурсів.

В таблиці 3 представлені зведені результати проведених експериментів. В останньому рядку таблиці підрахована кількість разів, коли визначений спосіб давав найкращі результати ранжування.

Таблиця 3

Підсумкові результати експериментів №1 - №4.

Номер експерименту	**Спосіб розрахунку значень за методом середніх балів**			
	СА	**САЗ**	**СГ**	**СГЗ**
№1. Сума відхилень:	26	**18**	20	**18**
Середнє значення відхилення:	1.3	**0,9**	1	**0,9**
№2. Сума відхилень:	32	28	32	**24**
Середнє значення відхилення:	1.6	1.4	1.6	**1.2**
№3. Сума відхилень:	32	30	24	**22**
Середнє значення відхилення:	1.6	1.5	1.2	**1.1**
№4. Сума відхилень:	22	22	**20**	**20**
Середнє значення відхилення:	1.1	1.1	**1**	**1**
Кількість кращих результатів:	**0**	**1**	**1**	**4**

З таблиці видно, що в ході проведення чотирьох експериментів, найменше відхилення від еталонного ранжування показав метод середніх балів, що розраховувались на основі середнього гармонічного зваженого.

У шостому розділі розглядається структура програмного комплексу, що реалізує всі вимоги до системи КОДІ, сформульовані в другому розділі.

Даний програмний комплекс розрахований на використання в режимі реального часу і побудований за технологією клієнт-сервер. Має простий, інтуїтивно зрозумілий інтерфейс для взаємодії користувача та адміністратора системи з кожним із представлених модулів. Для зберігання результатів роботи використовуються два основних формати: бази даних та система контролю версій.

Для побудови програмного комплексу, як і проектування системи КОДІ використовувалась гнучка модульна архітектура. В рамках цієї архітектури були програмно реалізовані всі описані в другому розділі модулі системи.

Доступ до програми реалізований через веб-інтерфейс. Для цього використовується perl модуль main.cgi на стороні сервера і модуль index.htm на стороні клієнта, який використовує javascript.

Окремо слід розглянути два основних модулі системи, це система керування вмістом для створення корпоративних веб-ресурсів з інтегрованою семантичною розміткою (рис. 6), та модуль інформаційного пошуку.

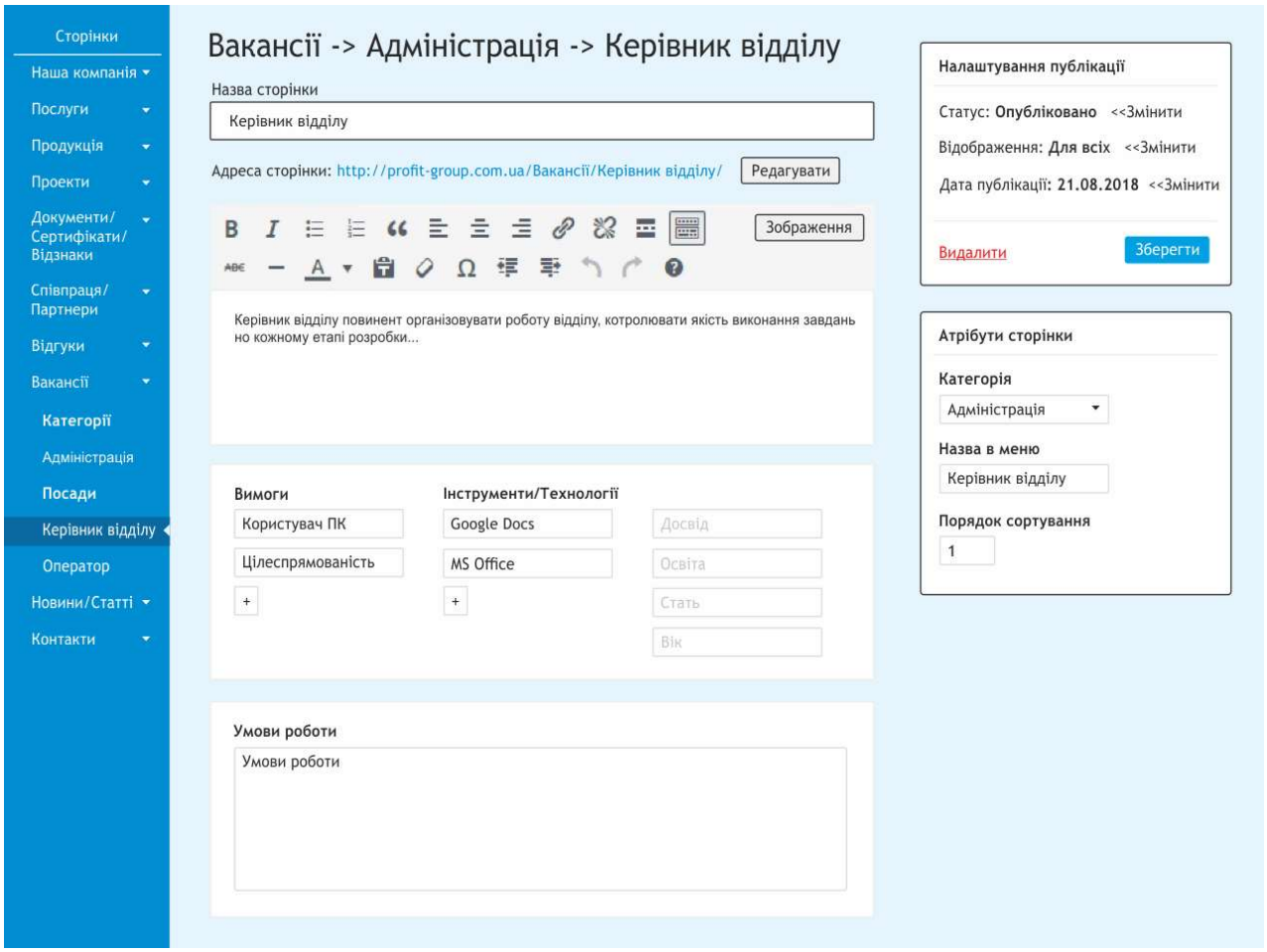


Рис. 6. Інтерфейс додавання вакансії

Система керування вмістом побудована на основі архітектури сучасних CMS і вміщує всі стандартні функції, такі як додавання сторінок, керування навігаційним меню, загальні налаштування системи, керування відгуками, тощо.

Розроблена CMS включає перелік типових сторінок. Для кожної сторінки реалізовано інтерфейс додавання специфічних для неї полів згідно розробленої структури корпоративних веб-ресурсів.

Модуль інформаційного пошуку реалізовано у вигляді метапошукової системи, що в якості базових результатів використовує пошукову видачу Google.com.ua.

В системі реалізовано ряд нових можливостей 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;
- розробка власних пошукових агентів, та їх застосування для додаткової обробки пошукової видачі;

На рис. 7 представлено пошуковий інтерфейс.

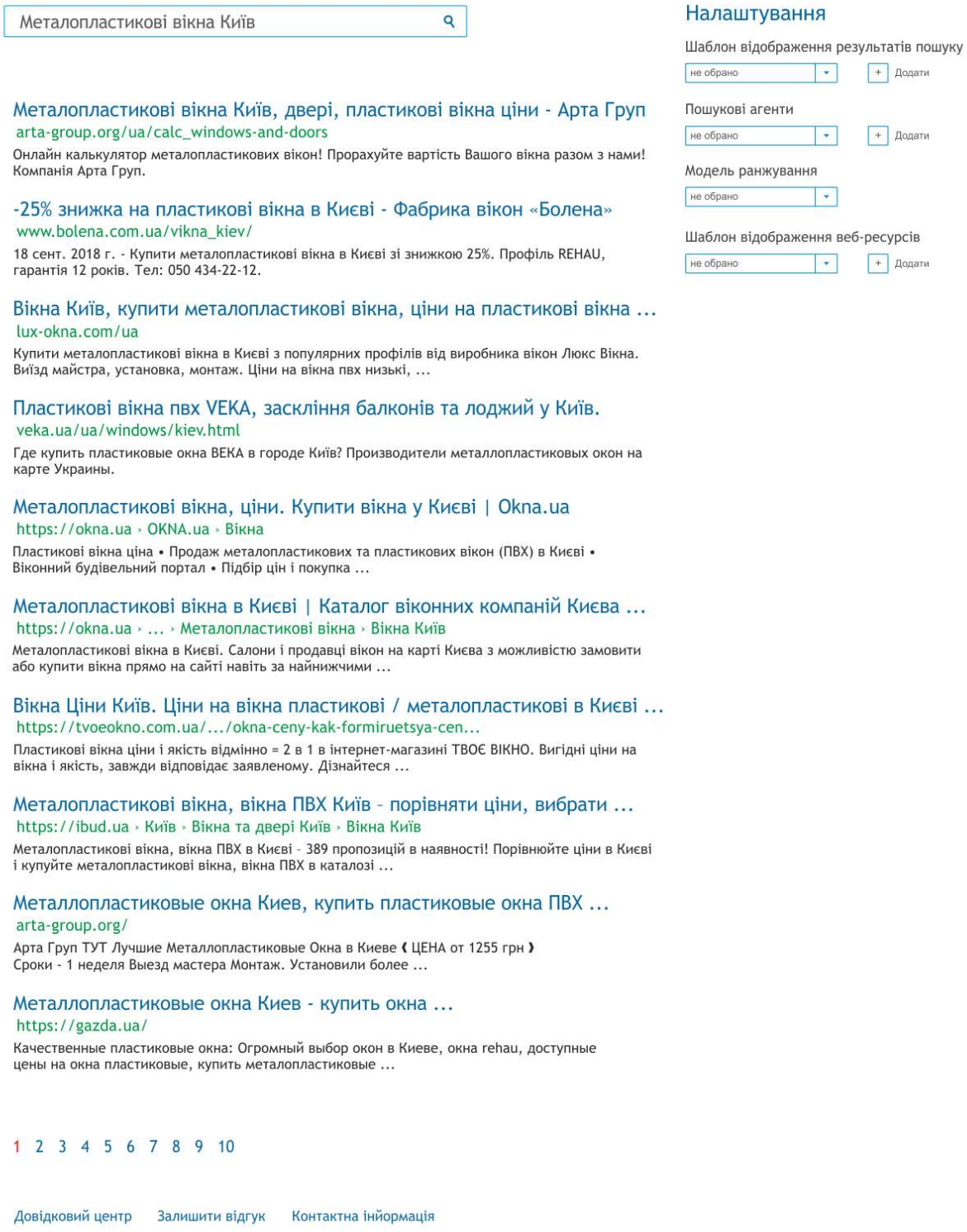


Рис. 7. Інтерфейс розробленої метапошукової системи

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

На рис. 8 зображено результати пошуку за запитом «металопластикові вікна Київ» із застосуванням шаблону «Текст+контакти» та пошукового агенту, що відбирає в пошуковій видачі Google всі компанії, що є офіційними дилерами або виробниками і в яких є відкриті вакансії.

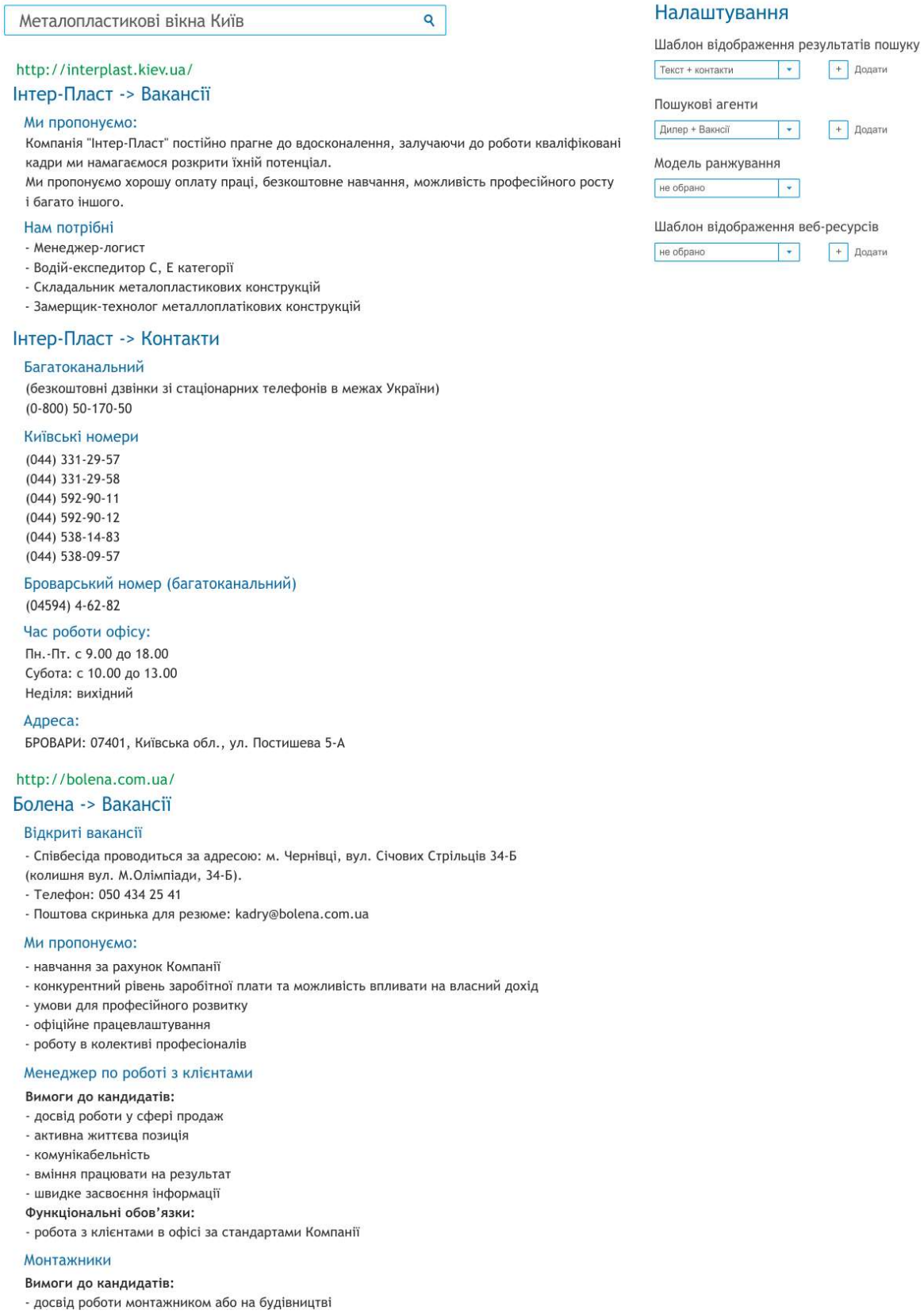


Рис. 8. Результати обробки пошукових даних

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

Висновки

На основі виконаних теоретичних та експериментальних досліджень у дисертаційній роботі вирішено важливу науково-прикладну проблему підвищення якості інтелектуального пошуку та обробки даних в мережі Інтернет із застосуванням програмних пошукових агентів, шляхом розробки комплексного підходу що об'єднує нові методи, моделі та засоби в єдину систему комплексного оперування веб-даними на семантичному рівні в режимі реального часу. Ефективність розроблених системи комплексного оперування даними в мережі Інтернет та програмного комплексу підтверджено експериментально та розв'язанням реальних прикладних задач.

Основні висновки та результати, отримані автором та представлені в дисертаційній роботі:

1. На основі аналізу інформаційного вмісту та елементів навігації 1000 корпоративних веб-ресурсів з вітчизняного сегменту всесвітньої павутини, побудовано їх загальну структуру та семантичний профіль що являє собою онтологію семантичних класів та властивостей для опису їх інформаційного наповнення.

2. Розроблено нову предметно-орієнтовану мову оперування веб-даними з широким функціоналом для видобування, обробки, збереження та відображення веб-даних. Всі можливості обробки інформації реалізовані в якості вбудованих засобів мови і орієнтовані в першу чергу на роботу з об'єктною моделлю документів, представлених у всесвітній павутині у вигляді HTML-коду. Основною перевагою даної мови є те, що вона реалізована в декларативному стилі і дозволяє розробнику описувати складні операції з даними простими вбудованими функціями. Всі можливості мови реалізовано з високим рівнем абстракції, що дозволяє змінювати окремі внутрішні механізми обробки даних, таких як формат збереження даних, методи розбору структури веб-сторінок, без необхідності корегування програмного коду. Розроблена мова стала ядром системи підвищення ефективності обробки веб-даних, та базою для реалізації платформи створення пошукових агентів, заснованої на семантичній розмітці веб-ресурсів.

3. Запропоновано новий підхід до унікальної ідентифікації фрагментів даних веб-сторінок в рамках доменного імені, заснований на використанні семантичних атрибутів HTML-блоків замість URL-адреси батьківської веб-сторінки, що дає можливість формувати шаблони відображення інформаційного вмісту веб-ресурсу з довільним розташуванням блоків даних не зважаючи на їх приналежність до певних сторінок веб-ресурсу.

4. Використовуючи новий підхід до унікальної ідентифікації фрагментів даних в рамках доменного імені розроблений новий метод відображення інформаційного вмісту веб-ресурсів на основі користувацьких шаблонів, що передбачає створення на стороні користувача одного шаблону з визначеною структурою розташування блоків даних та їх стильовим оформленням для відображення всіх веб-ресурсів певного типу. Може ефективно використовуватись для відсіву певних блоків веб-сторінок, що містять

нерелевантну інформаційним потребам користувача інформацію, а також для формування шаблонів відображення результатів пошукової видачі.

5. Розроблена нова персоналізована модель ранжування результатів пошуку інформації в мережі Інтернет на основі користувацьких оцінок. Модель базується на новому методі формування унікальних для кожного користувача експертних груп, заснованого на розрахунку міри узгодженості думок поточного користувача з потенційними експертами в режимі реального часу. Відповідно, для кожного користувача результати ранжування є унікальними, що робить цю модель набагато більш стійкою до впливу методів штучного підвищення рейтингу веб-ресурсів та пошукового спаму.

6. Розроблена абстрактна модель системи комплексного оперування даними в мережі Інтернет, дозволила об'єднати всі представлені в роботі підходи, методи, моделі та засоби в якості незалежних модулів, для забезпечення гнучкості та масштабованості системи. Висока ефективність системи, продемонстрована розв'язанням реальних прикладних задач, досягається за рахунок побудови ядра системи на основі розробленої предметно-орієнтованої мови оперування веб-даними, яка реалізує набір базових функцій обробки та перетворення даних.

7. Удосконалені методи інтеграції семантичної розмітки в HTML-код вже існуючих та нових веб-сторінок дозволяють більш ефективно використовувати всі переваги концепції семантичної павутини для видобування, обробки та аналізу веб-даних, а також розширюють можливості розробки та застосування методів інтелектуального пошуку.

8. Удосконалено технологію пошуку інформації в мережі Інтернет по ключових словах за рахунок розробки метапошукової системи, яка реалізує окремий рівень абстракції для додаткової обробки результатів роботи сучасних пошукових систем, що включає новий підхід до відображення результатів пошуку із застосуванням користувацьких шаблонів, нову персоналізовану модель ранжування результатів, а також можливість створення та застосування пошукових агентів безпосередньо в процесі пошуку.

9. На основі розробленої абстрактної моделі системи комплексного оперування даними в мережі Інтернет, розроблено програмний комплекс, з гнучкою модульною архітектурою, а також широкими можливостями налаштувань з боку користувача для всіх його модулів. Модулі системи поділяються за своїм призначенням на базові та функціональні. Базові модулі наслідують необхідні функції ядра системи і додають рівень абстракції між ядром та функціональними модулями. Функціональні модулі реалізують прикладні функції системи у вигляді веб-додатків для забезпечення інформаційних потреб кінцевого користувача. Кожен функціональний модуль має власні візуальний та текстовий інтерфейси для взаємодії клієнта з системою.

10. Програмна реалізація розробленої системи комплексного оперування даними в мережі Інтернет включає наступні функціональні модулі для взаємодії користувача з веб-даними:

- предметно-орієнтовану мову оперування веб-даними;

- метапошукову систему, засновану на додатковій обробці результатів роботи сучасних пошукових систем із застосуванням користувацьких шаблонів відображення результатів пошуку, нової моделі ранжування на основі оцінок користувачів, а також пошукових агентів;

- платформу розробки пошукових агентів на базі розробленої предметно орієнтованої мови оперування веб-даним та семантичного профілю корпоративних веб-ресурсів;

- платформу розробки шаблонів відображення вмісту веб-ресурсів на основі особистих потреб користувача;

- систему керування вмістом веб-ресурсів для створення веб-ресурсів з інтегрованою семантичною розміткою;

- інтерфейс інтеграції семантичної розмітки в HTML-код вже існуючих веб-сторінок.

11. Застосування програмного комплексу для розв'язання широкого кола реальних задач оперування веб-даними, показало його переваги перед існуючими методами та програмними рішеннями, що підтверджено актами впровадження результатів дисертаційного дослідження у виробничий процес вітчизняних підприємств та організацій.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Монографії та навчальні посібники:

1. Zosimov V., Stepashko V., Bulgakova O. Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results: monograph/ G.Setlak, M.Alexandrov, K.Markov (Eds.). Rzeszow, Poland; Sofia, Bulgaria: ITHEA, 2012. 345 p.

2. Зосімов В.В., Погромська Г.С., Махровська Н.А., Булгакова О.С. Інформаційний аналіз управління процесами в складних системах: теорія та практика: монографія. Миколаїв, МНУ ім. В.О.Сухомлинського, 2018. 217 с.

3. Степашко В.С., Зосімов В.В., Булгакова О.С. Ітераційні алгоритми індуктивного моделювання: монографія. Київ: Наукова думка, 2018. 190 с.

4. Булгакова О.С., Зосімов В.В., Погромська Г.С., Махровська Н.А. Інформатика: візуальне програмування: навч.-метод. посіб. Херсон: ОЛДІ-ПЛЮС, 2014. 312 с. (гриф МОН України, лист №1/11-1208 від 29.01.14)

Статті у наукових фахових виданнях України:

(які входять до переліку ВАК/ МОН України)

5. Дубовенко М.Н., Белов В.М., Зосимов В.В., Булгакова О.С. Гонтарь Т.М. Информационная технология анализа психологической информации из социальных сетей средствами индуктивного моделирования. Управляющие системы и машины, УСИМ. 2012. №5. С. 79-82, 92.

6. Зосимов В.В., Булгакова А.С., Поздеев В.А. Исследование технологии повышения эффективности поиска релевантной информации в Интернете. Вісник НУК, електронне видання. 2012. Вип. 5. URL: http://jn.nuos.edu.ua/ru/publication?publicationId=16471

7. Зосимов В.В., Поздеев В.А. Построение моделей ранжирования результатов поисковой выдачи для различных областей знаний. Вісник ЧДТУ, 2012. №4. С. 65-68.

8. Stepashko V., Bulgakova O. Zosimov V. Experimental Verification of Internal Convergence of Iterative GMDH Algorithms. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2012. Вип. 4. С. 38-42.

9. Зосимов В.В. Построение и применение модели отсеивания нерелевантных источников при поиске научно-технической информации в Интернете. Кибернетика и вычисл. техника. 2013. Вип. 171. С. 52-67.

10. Зосимов В.В., Булгакова А.С., Тищенко А.В. Пути и методы повышения быстродействия сложных вычислений. Комп'ютерні засоби, мережі та системи: Зб. наук. пр. Київ, 2013. Вип. 12. С. 27-33.

11. Zosimov V. Usage the operating systems multitasking for organization multi-access to complex systems modeling computing system. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2013. Вип. 5. С. 146-150.

12. Зосимов В.В. Планировщики задач операционной системи Linux и перспективи их развития для эффективного ведення научных расчетов. Проблеми моделювання: Зб. наук. пр. Київ, 2013. Вип. 70. С. 73-78.

13. Zosimov V. Prospects for Applying the Concept of the Semantic Web Analysis for Existing sites. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2014. Вип. 6. С. 41-46.

14. Зосимов В.В. Применение кластерного анализа для повышения полноты поиска информации в сети Интернет. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2015. Вип. 7. С. 123-128.

15. Зосимов В.В. Исследование значимости признаков ранжирования результатов работы поисковых систем. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2016. Вип. 8. С. 111-116.

16. Зосімов В.В., Ісмаїлов Ріфат Оріф огли, Свідоцтво про реєстрацію авторського права на твір "Комп'ютерна програма віддаленого керування електронними пристроями для бездротових команд" №69350 від 22.12.2016

17. Зосімов В.В. Застосування кластерного аналізу для відсіювання пошукового спаму. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2017. Вип. 9. С. 95-100.

18. Зосімов В.В., Христодоров О.В., Булгакова О.С. Технологія автоматизованої адаптації веб-додатків на основі ідентифікації кіберсутностей. Управляющие системы и машины, УСИМ. 2018. №3. С. 51-59.

19. Зосимов В.В. Комплексний підхід до підвищення якості пошуку інформації в мережі Інтернет. Управляющие системы и машины, УСИМ. 2018. №4. С. 32-45.

20. Зосимов В.В. Побудова особисто-орієнтованих моделей ранжування результатів пошуку інформації. Індуктивне моделювання складних систем: Зб. наук. пр. Київ, 2018. Вип. 9. С. 87-93.

21. Зосімов В.В. Свідоцтво про реєстрацію авторського права на твір "Система псевдоідентифікації користувачів на основі поведінкових характеристик" №83716 від 14.12.2018.

Статті у наукових фахових виданнях України,
які входять до міжнародних наукометричних баз даних:
(за умови, що на момент публікації статті видання вже внесене до міжнародної наукометричної бази даних)

22. Булгакова О.С., Зосімов В.В., Степашко В.С. Програмний комплекс моделювання складних систем на основі ітераційних алгоритмів МГУА з можливістю мережевого доступу. Системні дослідження та інформаційні технології. 2014. № 1. С. 43-55. (Index Copernicus)

23. Зосімов В.В., Горішня Д.О. Альтернативні види пошуку науково-технічної інформації в Інтернеті. Геометричне моделювання та інформаційні технології: науковий журнал. Миколаїв, 2016. №1. С. 69-73. (Index Copernicus)

24. Зосімов В.В., Саченко С.Г. Аналіз якості інформаційного наповнення сайту. Геометричне моделювання та інформаційні технології: науковий журнал. Миколаїв, 2018. № 1(5). С. 58 – 64. (Index Copernicus)

Статті в іноземних виданнях:
(тільки закордонні періодичні(ISSN) наукові фахові видання;
статті у міжнародних збірниках за матеріалами конференцій, навіть внесені до SCOPUS чи ін. міжнародної наукометричної бази даних сюди не входять)

25. Stepashko V., Bulgakova O., Zosimov V. Construction and research of the generalized iterative GMDH algorithm with active neurons. Advances in Intelligent Systems and Computing. Springer Verlag. 2018. Vol. 689. P. 492-510. ISSN 2194-5365

26. Zosimov V. Khrystodorov O., Bulgakova O. Dynamically changing user interfaces: software solutions based on automatically collected user information. Programming and Computer Software. 2018. Vol 44 (6). P. 492-498. Print ISSN 0361-7688

27. Зосимов В.В., Христодоров А.В., Булгакова А.С. Программные решения для динамического изменения пользовательского интерфейса на основе автоматически собранной информации о пользователе. Труды Института системного программирования. 2018. том 30 (вып. 3), С. 207-220. Print ISSN 2079-8156

28. 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 / Ed. by Markus Spies at al. Los Alamitos: IEEE Computer Society, 2015. P. 291-295 Print ISSN: 1529-4188

Статті в інших виданнях:
(статті у неперіодичних збірниках наукових праць (ISBN), у т.ч. статті у збірниках за матеріалами конференцій)

29. 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. P. 99-112. ISBN: 978-954-16-0057-3 (printed)

30. Stepashko V., Bulgakova O., Zosimov V. Numerical study of the generalized iterative algorithm GIA GMDH with active neurons. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, IEEE, Lviv, 6 November 2017. P. 496-500. ISBN: 978-1-5386-6463-6 (printed)

31. 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, Lviv, September 2018. P. 466-470 ISBN: 978-1-5386-6463-6 (printed)

Тези наукових доповідей:

32. Viacheslav Z., Stepashko V., Bulgakova O. Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results. Intelligent Information and Engineering Systems INFOS: proceedings of the 5th International Conference, 17th-21st September, Krynica-Zdrój, 2012. Rzeszow – Krynica-Zdrój, Poland. 2012. P. 196-201

33. Зосімов В.В., Булгакова О.С., Степашко В.С. Програмний комплекс моделювання складних систем на основі ітераційних алгоритмів МГУА з можливістю мережевого доступу. Системний аналіз та інформаційні технології, SAIT: матеріали 14 міжнар. наук.-техн. конф., м. Київ, 24 квітня 2012 р. Київ, 2012. С. 176-177

34. Зосимов В.В. Повышение эффективности информационного поиска с применением индуктивных алгоритмов. Информационные и коммуникационные технологии в науке, образовании и производстве: материалы междунар. научно-практ. конф., г. Тирасполь, 29-30 октяб. 2012 г. Молдова, 2012. С. 51-54

35. Зосимов В.В., Булгакова А.С. Использование GRID-систем для распределения процесса вычислений по данным в алгоритмах индуктивного моделирования. Інтелектуальні системи прийняття рішень та проблеми обчисленого інтелекту, ISDMCI: матеріали міжнар. наук. конф., м. Євпаторія, 20-24 травня 2013 р. Херсон, 2013. С. 448-450.

36. Zosimov V. Usage the Operating Systems Multitasking for Organization Multi-Access to Complex Systems Modeling Computing System. Inductive Modelling, ICIM: proceedings of the International Conference, on Kyiv, 21-24 July 2013, Kyiv, Ukraine 2013. P.53-57.

37. Зосимов В.В., Синевид К.М., Булгакова А.С. Извлечение и анализ данных Интернет-ресурсов с использованием семантической разметки. Компьютерное моделирование в наукоемких технологиях: материалы междунар. конф., г.Харьков, 28-31 мая, 2014 г. Харьков, 2014. С. 88-90.

38. Булгакова А.С., Зосимов В.В. Интеллектуальные методы анализа данных как инновационный элемент мониторинга экономических процессов. Геоинформационные системы и компьютерные технологии эколого-

экономического мониторинга: материалы междунар. научно-техн. конф., г. Днепропетровск, 9-11 апреля, 2014 г. Днепропетровск, 2014.

39. Булгакова А.С., Зосимов В.В. Интеллектуальный подход к извлечению знаний из данных с помощью алгоритмов индуктивного моделирования. Проблемы теоретической кибернетики: материалы междунар. конф. г.Казань, 16-20 июня, 2014 г. Казань, Россия, 2014. С. 47-50.

40. Zosimov V. Web-based information analysis using inductive modeling algorithms. Inductive Modelling: proceedings of the International Workshop, Zhukyn IWIM, 20-25 July, 2015. Kyiv, 2015. P. 81-84.

41. Зосимов В.В., Булгакова А.С., Емец М.М. Механизм извлечения и анализа данных слабоструктурированных веб-документов с использованием семантической разметки Інтелектуальні системи прийняття рішень та проблеми обчисленого інтелекту, ISDMCI: матеріали міжнар. наук. конф., м. Залізний порт, 24-28 травня 2016 р. Херсон, 2016. С. 118-120.

42. Zosimov V. The Significant Features for Search Engines Ranking Results. Inductive Modelling: proceedings of the International Workshop, Zhukyn IWIM, 18-20 July, 2016. Kyiv, 2016. P. 54-60.

43. Зосімов В.В., Белінський Д.А. Вплив сучасних тенденцій UX\UI дизайну на розвиток веб-ресурсів. Інформаційні технології в моделюванні: матеріали всеукраїнської наук.-практ. конф. студ., аспір. та молодих вчених, м.Миколаїв, 23-24 березня, 2017 р. Миколаїв, 2017. С.133-134.

АНОТАЦІЯ

Зосімов В.В. Моделі та засоби інтелектуальної обробки даних корпоративних веб-ресурсів. – Рукопис.

Дисертація на здобуття наукового ступеня доктора технічних наук зі спеціальности 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем. – Київський національний університет імені Тараса Шевченка Міністерства освіти і науки України. – Київ, 2019.

В дисертаційний роботі на основі виконаних теоретичних та експериментальних досліджень, вирішено важливу науково-прикладну проблему підвищення якості інтелектуального пошуку та обробки даних в мережі Інтернет із застосуванням програмних пошукових агентів, шляхом розробки комплексного підходу що об’єднує нові методи, моделі та засоби в єдину систему комплексного оперування веб-даними на семантичному рівні в режимі реального часу.

Запропонований підхід включає: загальну структуру та семантичний профіль корпоративних веб-ресурсів, нову предметно-орієнтовану мову оперування веб-даними з широким функціоналом для видобування, обробки, збереження та відображення веб-даних, новий підхід до унікальної ідентифікації фрагментів даних веб-сторінок в рамках доменного імені, новий метод відображення інформаційного вмісту веб-ресурсів на основі користувацьких шаблонів, нову персоналізовану модель ранжування результатів Інтернет-пошуку на основі оцінок користувачів.

Розроблені моделі, методи і засоби об'єднані в систему комплексного оперування даними в мережі Інтернет, а також втілені у вигляді програмного комплексу, що реалізує функціональні можливості системи.

Ключові слова: пошук інформації, ранжування, онтологія, семантичний профіль, веб-ресурс, аналіз даних, видобування інформації, пошукові агенти, предметно-орієнтована мова, метапошукова система, шаблони відображення даних.

АННОТАЦИЯ

Зосимов В.В. Модели и средства интеллектуальной обработки данных корпоративных веб-ресурсов. – Рукопись.

Диссертация на соискание ученой степени доктора технических наук по специальности 01.05.03 – математическое и программное обеспечение вычислительных машин и систем. - Киевский национальный университет имени Тараса Шевченко Министерства образования и науки Украины. – Киев, 2019.

В диссертационной работе на основе выполненных теоретических и экспериментальных исследований, решена важная научно-прикладная проблема повышения качества интеллектуального поиска и обработки данных в сети Интернет с применением программных поисковых агентов, путем разработки комплексного подхода, объединяющего новые методы, модели и средства в единую систему комплексного оперирования веб-данными на семантическом уровне в режиме реального времени.

Предложенный подход включает: общую структуру и семантический профиль корпоративных веб-ресурсов, новый предметно-ориентированный язык оперирования веб-данными с широким функционалом для извлечения, обработки, хранения и отображения веб-данных, новый подход к уникальной идентификации фрагментов данных веб-страниц в рамках доменного имени, новый метод отображения информационного содержания веб-ресурсов на основе пользовательских шаблонов, новую персонализированную модель ранжирования результатов Интернет-поиска на основе оценок пользователей.

Разработанные модели, методы и средства объединены в систему комплексного оперирования данными в сети Интернет, а также реализованы в виде программного комплекса.

Ключевые слова: поиск информации, ранжирование, онтология, семантический профиль, веб-ресурс, анализ данных, извлечение информации, поисковые агенты, предметно-ориентированный язык, метапоисковая система, шаблоны отображения данных.

ANNOTATION

Zosimov V.V. Models and tools of corporate web resources' intelligent data processing. – Manuscript.

Thesis for scientific degree of Doctor of Technical Sciences on the specialty 01.05.03 – mathematical and program software of computing machines and systems. –

Taras Shevchenko National University of Kyiv Ministry of education and science of Ukraine. – Kyiv, 2019.

The thesis, based on performed theoretical and experimental studies, is devoted to solving an important applied science problem of improving the quality of intellectual search and data processing on the Internet using the software search agents by developing a comprehensive approach that integrates new methods, models and tools into a single semantic-level system of complex web data operation in real time.

Based on the analysis of information content and navigation elements of a large number of corporate web resources, their general structure and semantic profile are constructed, that is a set of semantic classes and properties that describes their content.

A new approach is proposed to uniquely identify web page data snippets within a domain name, based on the use of semantic attributes of HTML blocks instead of the parent web page URL, it allows to form templates for displaying a web resource content with arbitrary arrangement of data blocks regardless of their belonging to certain pages of a web resource.

Using a new approach to the unique data fragments identification within a domain name, a new method was developed - displaying content information on web resources based on custom templates, it involves the creation of a single user-defined template with a defined structure of data blocks and their styling to display all web resources of a certain type. It can be effectively used for weeding out certain blocks of web pages that contain information irrelevant to the user's information needs, as well as to form patterns for displaying search results.

A new personalized model of ranking of search results on the Internet, based on user ratings has been developed. The model is based on a new method that is forming unique to each user expert group based on calculating the degree of consistency of the current user's opinions with potential experts in real time. Accordingly, the ranking results are unique to each user, which makes this model much more resilient to the impact of artificial web rank and search spam techniques.

The developed abstract model of the system of complex data management on the Internet allowed to combine all approaches, methods, models and tools presented in the work as independent modules, in order to ensure the flexibility and scalability of the system. High system performance, demonstrated by solving real-world application problems, is achieved by building the system kernel based on object-oriented web data management language that implements a set of basic data processing and conversion functions.

Improved methods for integrating semantic markup into the HTML code of existing and new web pages make more effective use of the semantic web concept benefits for web data extracting, processing and analyzing, as well as expanding the development and application of intellectual search methods.

The technology of searching information on the Internet by keywords has been improved by developing metasearch engine system which implements separate abstraction level for additional processing results of the modern search engines, it includes a new approach for the search results displaying using custom templates, a

new personalized results ranking model, as well as the ability to create and use search agents directly in the search process.

On the basis of the developed abstract model of system of complex data operation on the Internet a software module with flexible modular architecture was developed, with flexible modular architecture and extensive user-configurable features for all its modules. System modules are divided by their purpose into basic and functional. Basic modules inherit the necessary functions of the kernel of the system and add a level of abstraction between the kernel and the functional modules. Functional modules implement the application functions of the system in the form of web applications to meet the information needs of the end user. Each functional module has its own visual and textual interfaces for client interaction with the system.

Application of the software complex for solving a wide range of real-world problems of web data management has shown its advantages over existing methods and software solutions, which is confirmed by the acts of introduction of the dissertation results into the production process of domestic enterprises and organizations.

Keywords: information search, ranking, ontology, semantic profile, web resource, data analysis, information retrieval, search agents, domain-specific language, metasearch system, data display templates.