

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

ПОЧАНСЬКИЙ ОЛЕГ МИХАЙЛОВИЧ

УДК 004.853

**МЕТОДИ СИНТЕЗУ АДАПТИВНИХ WEB-СТОРІНОК НА ОСНОВІ
ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ІНФОРМАЦІЙНИХ РЕСУРСІВ МЕРЕЖІ
ІНТЕРНЕТ**

05.13.23 – системи та засоби штучного інтелекту

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2013

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Терзіян Ваган Якович,
Харківський національний університет
радіоелектроніки, професор кафедри
штучного інтелекту.

Офіційні опоненти: доктор технічних наук, доцент
Литвин Василь Володимирович,
Національний університет «Львівська
політехніка», доцент кафедри
«Інформаційні системи та технології»,
м. Львів.

доктор технічних наук, доцент
Шостак Ігор Володимирович,
Національний аерокосмічний
університет ім. Н.С. Жуковського
«ХАІ», професор кафедри інженерії
програмного забезпечення, м. Харків.

Захист відбудеться « 19 » червня 2013 р. о 13-00 годині на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитися у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий « 17 » травня 2013 р.

Учений секретар
спеціалізованої вченої ради

Є.І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Інтенсивний розвиток Інтернет-технологій останнім часом призводить до того, що тепер кожен користувач глобальної мережі може отримати доступ до практично необмежених інформаційних ресурсів. З появою концепції «Web 2.0» інтернет-користувач із звичайного отримувача інформації перетворився на активного учасника інформаційного процесу, який вже сам може створювати власні, редагувати та коментувати інші електронні документи. За даними української дослідницької компанії УНІАН GFK Ukraine чисельність регулярних інтернет-користувачів в Україні зросла у 2012 році до 15 мільйонів осіб. Така популярність використання Web-інформації викликала значне збільшення серйозних наукових досліджень, які пов'язані з розвитком пошукових сервісів та розробкою методів підвищення релевантності (відповідності запитам) даних, отриманих користувачами.

Це пояснюється, зокрема, появою великої кількості тематичних Web-сторінок, які здебільшого містять низькоякісні або неперевірені дані або просто дублюють одна одну, тобто є нерелевантними джерелами інформації. Крім того, навіть релевантні електронні документи можуть бути зашумленими через присутність реклами або непотрібної інформації, що не відповідає їх загальній тематиці. Із збільшенням рівня таких шумів помітно знижується ефективність сприйняття підсумкового тексту кінцевим користувачем.

У зв'язку з цим, у сучасних умовах, щоб знайти релевантну інформацію, яка відповідає критеріям користувача, необхідно витратити чимало часу на обробку різноманітних джерел за тематикою, яка його цікавить. Нажаль, суттєво не покращують ситуацію і пошукові системи, які іноді видають за одним запитом тисячі релевантних результатів, що також не сприяє підвищенню ефективності пошуку (під ефективністю пошуку зазвичай розуміють знаходження необхідної та достовірної інформації за мінімально можливий час). З іншого боку, останнім часом активно розвиваються спеціалізовані системи видобування даних із Web-джерел. Їх основна особливість полягає в структурованості оброблюваної інформації за ключовими атрибутами, що дозволяє у загальному випадку підвищити ефективність пошуку релевантних даних за запитами користувачів. Але більшість існуючих методів такої структуризації має ряд суттєвих обмежень та припущень, пов'язаних зі складністю алгоритмів аналізу та обробки необхідних даних за допомогою спеціалізованих пошукових систем. В першу чергу це пов'язано зі складностями розуміння пошуковою системою сутності значущої інформації у її лексичному сенсі та її відміну від іншої інформації, що не вдовольняє потребам або критеріям, які поставлені перед даною системою.

Рішенню проблем, які при цьому виникають, зокрема сприяє поява нових стандартів та технологій, розроблених консорціумом World Wide Web Consortium (W3C) для опису електронних Web-документів в рамках аналізу інформаційних ресурсів мережі Інтернет. Зокрема, для опису соціального профілю користувача використовують стандарти FOAF и Dublin Core, а для

опису електронних інтернет-ресурсів – стандарт SPARQL. Цей же стандарт використовується і для створення запитів к онтологічним базам знань, які є частиною технології Semantic Web. Останнім часом з'явилися чисельні наукові публікації, присвячені теоретичним та експериментальним дослідженням в галузі розробки методів створення онтологій електронних документів та інтелектуалізації пошуку релевантних Web-даних з використанням вищезгаданих стандартів і технологій. Суттєвий вклад у розвиток даного напрямку штучного інтелекту внесли, зокрема, Berners-Lee T., Bergman M., Hendler J., Manning C., Perkowski M., Valasquez J., Агеев М. С., Гаврилова Т.А., Зеленков П.В., Карцан І.Н., Ланде Д.В., Леонтьєва Н.Н., Терзіян В.Я., Хорошевський В.Ф. та інші вчені.

Разом з цим недостатньо дослідженими є задачі розвитку пошукових систем та методів, пов'язаних з необхідністю перевірки на зашумленість і повторюваність інформації у досліджуваних документах; з відсутністю критеріїв семантичного пошуку, які повною мірою враховують соціальний профіль користувача та можливість його динамічного коригування; з відсутністю ефективних алгоритмів попередньої структуризації Web-даних, які аналізуються пошуковими системами; а також з відсутністю методів створення персоналізованих інтернет-ресурсів, які складаються з адаптивних Web-сторінок.

Таким чином, актуальною та важливою науково-технічною задачею є розробка методів синтезу адаптивних методів та відповідних програмних засобів для підвищення ефективності роботи системи семантичного пошуку та аналізу даних у мережі Інтернет на основі використання нових критеріїв оцінювання їх релевантності. Задачі, які при цьому виникають, зумовили напрямок досліджень даної дисертаційної роботи.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційну роботу виконано згідно з планом науково-дослідної роботи ХНУРЕ в рамках держбюджетних тем № 233 «Розробка системи підтримки семантичних запитів до онтологічної бази акредитації та ліцензування» (№ ГР 0109U001647) та № 243-4 «Методи створення загальної онтологічної бази соціально-економічної освітньо-наукової мережі з метою її інтеграції у європейський простір» (№ ГР 0109U002497), у яких здобувач приймав участь як виконавець. В рамках вказаних тем здобувачем запропоновані методи побудови та програмні агенти інтелектуальної системи семантичного пошуку релевантної інформації та синтезу адаптивних Web-сторінок.

Мета і задачі дослідження.

Мета дослідження – розробка методів синтезу адаптивних Web-сторінок для підвищення ефективності роботи експертної системи семантичного пошуку та аналізу даних у мережі Інтернет на основі використання нових критеріїв оцінювання їх релевантності. Для досягнення цієї мети у роботі вирішені такі основні задачі:

1. Дослідження задачі семантичного пошуку даних у мережі Інтернет та інтелектуального аналізу Web-ресурсів.

2. Розробка методу інтелектуального пошуку та аналізу інформації у мережі Інтернет, ґрунтованого на використанні нових критеріїв, які дозволяють враховувати значущість Web-ресурсів для кожного користувача залежно від його соціального профілю та оцінювати рівень їх зашумленості.

3. Розробка методу визначення структурних характеристик Web-документів на основі використання кластерного аналізу та нечіткої логіки, який дозволяє кількісно оцінювати ступінь привабливості аналізованих електронних даних для користувача.

4. Вдосконалення методу структуризації інформації на основі онтологічного підходу та технології Semantic Web з метою створення та обробки семантичних описів Web-ресурсів для експертної системи семантичного пошуку та аналізу даних у мережі Інтернет.

5. Розробка методу формування адаптивних Web-сторінок з використанням універсальної програми-агента, що забезпечує оперативний доступ користувача до інформації в Інтернет, яка його цікавить.

6. Програмна реалізація та тестування системи семантичного пошуку і аналізу даних в мережі Інтернет, що базується на використанні отриманих в дисертації теоретичних результатів.

Об'єкт дослідження – семантичний пошук та структурування інформації у Web-просторі на основі аналізу соціального профілю користувача.

Предмет дослідження – методи синтезу адаптивних Web-сторінок для підвищення ефективності роботи експертної системи семантичного пошуку та аналізу даних у мережі Інтернет на основі використання нових критеріїв оцінювання їх релевантності та стандартів Semantic Web.

Методи дослідження. Для вирішення сформульованих в роботі задач застосовувались такі методи: методи семантичного пошуку інформації – для розробки алгоритмів пошуку та аналізу Web-документів з використанням запропонованих в дисертаційній роботі критеріїв; методи кластерного аналізу та нечіткої логіки – для формування та кількісного оцінювання структурних характеристик Web-документів; методи агентних технологій – для формування адаптивних Web-сторінок з використанням універсального програми-агента; методи формування онтологічних моделей – для структуризації інформації, яка цікавить користувача, та семантичного опису Web-ресурсів; методи проектування складних Web-додатків – для програмної реалізації експертної системи семантичного пошуку та аналізу даних в мережі Інтернет, яка базується на використанні отриманих в дисертації теоретичних результатів.

Наукова новизна отриманих результатів.

Наукова новизна результатів, отриманих під час виконання дисертаційної роботи, полягає в наступному:

1. Вперше запропоновано метод інтелектуального пошуку та аналізу інформації в мережі Інтернет, що базується на використанні нових критеріїв, які дозволяють враховувати значущість Web-ресурсів для кожного користувача залежно від його соціального профілю та оцінювати рівень їх зашумленості.

2. Вперше запропоновано метод визначення структурних характеристик Web-документів на основі використання кластерного аналізу та нечіткої логіки, який дозволяє кількісно оцінювати ступінь привабливості для користувачів даних, що аналізуються.

3. Вдосконалено метод структуризації інформації на основі онтологічного підходу та технології Semantic Web. На відміну від існуючих методів він дозволяє створювати та обробляти семантичні описи Web-ресурсів для підвищення ефективності системи семантичного пошуку та аналізу даних.

4. Отримав подальший розвиток метод формування адаптивних Web-сторінок шляхом використання універсальної програми-агента, яка забезпечує оперативний доступ користувача до цікавої для нього інформації в мережі Інтернет. Його відмінною особливістю є можливість накопичування та оперативного надання тематичної інформації згідно з поточним запитом користувача та його соціальним профілем.

Практичне значення отриманих результатів полягає в тому, що розроблені критерії та методи можуть бути використані під час створення та практичної реалізації інтелектуальних систем пошуку інформації в мережі Інтернет з урахуванням соціального профілю користувачів.

За допомогою розроблених методів та спеціалізованих програмних продуктів найбільш ефективно можна вирішувати такі практичні задачі:

- проектування систем пошуку релевантних Web-документів, які враховують семантику шуканої інформації та інтереси користувачів;
- модифікація існуючих інформаційно-пошукових систем з метою підвищення релевантності даних, отриманих з мережі Інтернет за запитом користувачів;
- проектування соціально-направлених систем для обміну інформацією між користувачами з близькими інтересами;
- розробка адаптивних Web-сторінок на основі даних, отриманих з мережі Інтернет, з можливістю коригування інтересів та соціального індексу користувачів;
- формування та оперативне коригування бази даних інформаційно-пошукових систем на основі індексування та виділення значущої інформації із різних Web-ресурсів;
- тематико-орієнтований моніторинг користувачів мережі Інтернет.

На основі запропонованих методів та критеріїв пошуку та аналізу інформації в мережі Інтернет розроблено та доведено до реалізації спеціалізований Web-ресурс, основними задачами якого є пошук електронних документів, які задовольняють актуальним запитам користувачів, та побудова інтерактивних Web-сторінок з використанням обраних джерел в рамках заданої тематики.

Результати дисертаційної роботи використані: для розробки експертно-пошукової системи газети «Пожежний вісник» (акт впровадження від 04.10.2012р.); у навчальному процесі під час підготовки та модифікації курсів «Методи видобування у Web-просторі», «Web-сервіси та сервіс-орієнтовані

архітектури», «Експертні системи, «Web-технології» на кафедрі штучного інтелекту Харківського національного університету радіоелектроніки (акт впровадження від 28.09.2012р.).

Особистий внесок здобувача. Всі наукові положення та практичні рекомендації отримані автором самостійно та опубліковані у 16 одноосібних наукових роботах, в яких відображені наступні результати: у [1-4] розглянуті основні принципи побудови семантичних анотацій складно структурованих компонентів Web 2.0 (блогів і форумів); у [5, 6] запропоновано модель побудови адаптивних Web-сторінок на основі інтелектуального аналізу контенту мережі Інтернет; у [7] розроблено математичну модель виділення структурних характеристик Web-документа та привласнення їм різних числових коефіцієнтів з використанням кластерного аналізу та нечіткої логіки; у [1, 8, 9] запропоновано модель виділення та оцінювання значущої інформації із Web-документів з урахуванням соціального індексу користувачів; у [10-16] запропоновано та досліджено метод побудови спеціалізованого Web-ресурса, основними задачами якого є пошук електронних документів, які задовольняють актуальним запитам користувачів, та побудову інтерактивних тематичних Web-сторінок.

Апробація результатів роботи. Основні положення та результати дисертаційної роботи доповідались та обговорювались на 13-м та 16-м Міжнародних молодіжних форумах «Радіоелектроніка і молодь в 21 сторіччі» (Харків, 2009, 2012); на 1-й науково-практичній конференції «Інтелектуальні системи – Біоніка інтелекту» (Харків, 2008); на 5-й та 8-й міжнародних науково-практичних конференціях «Математичне та програмне забезпечення інтелектуальних систем» (Дніпропетровськ, 2008, 2010); на 1-й науково-практичній конференції «Сучасні напрямки розвитку інформаційно-комунікаційних технологій та засобів керування» (Харків – Київ, 2010); міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (Євпаторія, 2011); 1-й науково-технічній конференції «Обчислювальний інтелект (результати, проблеми, перспективи)» (Черкаси, 2011); на 6-й міжнародній науково-практичній конференції «Наука та соціальні проблеми суспільства: інформатизація та інформаційні технології» (Харків, 2011); на міжнародній науково-практичній конференції «Інформаційні технології та інформаційна безпека в науці, техніці та освіті» (Севастополь, 2011).

Публікації. За результатами досліджень опубліковано 16 одноосібних наукових робіт, з них: 6 статей в наукових виданнях, які входять у затверджений перелік наукових фахових видань України та 10 робіт у збірниках праць наукових конференцій та семінарів.

Структура дисертації. Дисертація складається із вступу, чотирьох розділів, висновків, списку використаних джерел та двох додатків. Повний обсяг дисертації становить 168 сторінок; обсяг основного тексту 147 сторінок; 42 ілюстрації; 3 таблиці; список використаних джерел, що включає 122 найменування та займає 13 сторінок; два додатки на 8 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми, сформульовано мету та задачі дослідження, наведено відомості про зв'язки вибраного напрямку дослідження із планами організації, де виконана робота. Дано стислу анотацію отриманих в дисертації результатів, зазначено їх практичну цінність, обґрунтованість і достовірність. Наведено дані про використання проведених досліджень в навчальному процесі та розробці електронних видань.

Перший розділ присвячено аналізу існуючих технологій та методів пошуку інформації у мережі Інтернет, що є релевантною до запитів користувачів. Показано, що недостатньо дослідженими є задачі розвитку пошукових систем та методів, пов'язаних з перевіркою на зашумленість та повтореність інформації в досліджуваних документах; з відсутністю критеріїв семантичного пошуку, що в повній мірі враховують соціальний профіль користувачів та можливість його динамічної корекції; з відсутністю ефективних алгоритмів структуризації Web-даних, що аналізуються пошуковими системами; а також з відсутністю методів створення персоналізованих Інтернет-ресурсів, що складаються з адаптивних Web-сторінок. Згідно з існуючою термінологією, під адаптивними Web-сторінками розуміють динамічні персоналізовані Web-сторінки, що автоматично змінюють свою структуру та зміст з урахуванням актуальних преференцій користувача та характеру його запитів до інформаційних ресурсів мережі Інтернет.

Таким чином, можна зробити висновок щодо доцільності розробки спеціалізованої гібридної пошукової системи з можливістю генерування адаптивних Web-сторінок, що має реалізовувати наступні функціональні можливості:

- брати до уваги основні інтереси користувача під час пошуку потрібної йому інформації;
- здійснювати ранжування знайденої інформації згідно з поточними інтересами користувача;
- оцінювати привабливість Web-ресурсу для користувача;
- працювати з різними ресурсами мережі Інтернет, доступ до яких має кожен з користувачів;
- реалізовувати основні переваги концепції Web 2.0 (можливість оцінювати та коментувати джерела інформації) та Web 3.0 (пошук джерел інформації на основі їх семантики);
- використовувати профіль користувача у соціальних мережах для автоматичного заповнювання реєстраційної форми;
- оцінювати ступень значущості кожного Web-документа для поточного користувача при формуванні підсумкового результату пошуку;
- складати максимально повний перелік структурних характеристик Web-документа будь якого формату;

- застосовувати складений перелік структурних характеристик Web-документів для прогнозування підсумкового результату пошуку потрібної користувачеві інформації;
- забезпечувати користувача інформацією з різних джерел мережі Інтернет у реальному часі;
- формувати ефективну базу даних з можливістю її постійного оновлення у вигляді онтології тематичних Web-ресурсів з урахуванням їх семантичних зв'язків.

Для реалізації цих функцій згідно з метою дисертаційного дослідження необхідно розробити методи синтезу адаптивних Web-сторінок, що передбачають побудову ефективних процедур семантичного пошуку та аналізу даних в мережі Інтернет на основі застосування нових критеріїв їх релевантності.

Другий розділ присвячено розробці методів пошуку та інтелектуального аналізу релевантності інформації з сумісним використанням соціального індексування та структурних характеристик Web-ресурсів та соціальних профілів користувачів. Зокрема, вперше запропоновано метод визначення структурних характеристик Web-документів на основі використання кластерного аналізу та нечіткої логіки, який дозволяє кількісно оцінювати ступінь привабливості для користувачів даних, що аналізуються.

Цей метод базується на припущенні щодо можливості розбиття сучасних Web-ресурсів на різні інформаційні блоки. Кожен з цих блоків, у свою чергу, містить релевантні дані, присвячені заданій тематиці (у ньому присутня значуща інформація), та характеризується сукупністю Html-тегів, що повторюються. Існує відповідний набір унікальних характеристик, які дозволяють формувати блоки зі значущою інформацією. Зрозуміло, що близькість між Web-документами (або їх інформаційними блоками) визначається в залежності від метричної відстані. Якщо кожний документ описується n властивостями (ознаками), то він може бути представлений точкою у n -вимірному просторі, а близькість з іншими документами буде визначатися як відповідна відстань.

Враховуючи вагові коефіцієнти, значення яких відповідають експертним оцінкам, меру близькості між об'єктами (Web-документами) можна представити у наступному вигляді:

$$d_{Eij} = \left(\sum_{l=1}^n \omega_l (x_i^l - x_j^l)^2 \right)^{\frac{1}{2}}, \quad (1)$$

де $l = \overline{1, n}$ (n – загальна кількість усіх ознак об'єктів); x_i^l, x_j^l – значення l -ї ознаки для i -го та j -го об'єктів відповідно; ω_l – вага l -ї ознаки.

Для зведення всіх величин до діапазону значень $[0, 1]$, необхідно здійснити нормалізацію значень усіх ознак об'єктів:

$$x(normal)_i^l = \frac{x_i^l - x(min)^l}{x(max)^l - x(min)^l}, \quad (2)$$

де $x(normal)_i^l$ – нормалізоване значення l -ї ознаки i -го об'єкту; $x(max)^l$, $x(min)^l$ – максимально та мінімально можливі значення l -ї ознаки відповідно.

Таким чином, формула визначення відстані між i -м та j -м Web-об'єктами набуває вигляду:

$$d_{Eij} = \left(\sum_{l=1}^n \omega_l \left(\frac{x_i^l - x_j^l}{x(max)^l - x(min)^l} \right)^2 \right)^{\frac{1}{2}}. \quad (3)$$

Для визначення відстані між інформаційними блоками та знаходження блоку зі значущою інформацією розглянемо набір наступних характеристик: кількість тегів, що відповідають за форматування тексту (наприклад, <p>, <h1>,
 і т.д.); кількість спеціальних символів для включення тексту до Web- сторінки (наприклад, & > <); кількість символів, що застосовуються в тегах форматування електронного документа.

Крім того, для обчислення точного значення евклідової відстані для кожного з блоків Web-ресурса необхідно визначити характеристики об'єктів, що мають фіксовані значення (насамперед, шаблон елемента Web-документа з найгіршими характеристиками). Таким чином, загальна формула визначення евклідової відстані між шаблоном w та кожним з блоків аналізованого Web-ресурса (з урахуванням нормалізації величин) набуває наступного вигляду:

$$d_{Eiw} = \sqrt{\omega_1 \left(\frac{Tg_w - Tg_i}{Tg(max) - Tg(min)} \right)^2 + \omega_2 \left(\frac{Sp_w - Sp_i}{Sp(max) - Sp(min)} \right)^2 + \omega_3 \left(\frac{Sm_w - Sm_i}{Sm(max) - Sm(min)} \right)^2}, \quad (4)$$

де i – номер блоку аналізованого Web-ресурса; w – шаблон елемента Web-документа з найгіршими характеристиками; d_{Ei} – евклідова відстань для i -го блоку аналізованого Web-ресурса; $\omega_{1,2,3}$ – вагові коефіцієнти для характеристик i -го блоку Web-ресурса (задаються експертами або визначаються емпірично); Tg – кількість тегів, що відповідають за форматування тексту; $Tg(max)$, $Tg(min)$ – максимально та мінімально можливі значення Tg серед усіх аналізованих Web-ресурсів відповідно; Sp – кількість спеціальних символів для включення тексту до Web- сторінки; $Sp(max)$, $Sp(min)$ – максимально та мінімально можливі значення Sp серед усіх аналізованих Web-ресурсів відповідно; Sm – кількість символів, що застосовуються в тегах форматування електронного документа; $Sm(max)$, $Sm(min)$ – максимально та мінімально можливі значення Sm серед усіх аналізованих Web-ресурсів відповідно.

З урахуванням того, що для шаблона w значення характеристик мають бути мінімальними, формулу (4) можна представити у наступному вигляді:

$$d_{E_i} = \sqrt{\omega_1 \left(\frac{Tg_i}{Tg_i(max)} \right)^2 + \omega_2 \left(\frac{Sp_i}{Sp_i(max)} \right)^2 + \omega_3 \left(\frac{Sm_i}{Sp_i(max)} \right)^2}. \quad (5)$$

Таким чином, блок аналізованого Web-ресурсу з максимальною евклідовою відстанню до шаблону, визначається як блок зі значущою інформацією. Після формування блоків зі значущою інформацією необхідно визначити найбільш якісний Web-документ з найменшим рівнем шумів. Для оцінки зашумленості джерел інформації аналізуються блоки з незначущою інформацією Web-ресурса. Як еталонний об'єкт у роботі використано шаблон, що об'єднує п'ять блоків, один з яких забезпечує розміщення значущої інформації, а інші формують зовнішній дизайн.

При розробці метода формування та оцінювання структурних характеристик Web-документів були прийняті до уваги такі характеристики як розмір $T(Size)$, час та дата створення $T(Date)$, об'єм тексту $T(Capacity)$, зашумленість (відсоткове відношення шумів до всього тексту на сторінці) $T(Noisiness)$, кількість рисунків (ілюстрацій, графіків, фотографій) в тексті $T(Pictures)$, кількість посилань на зовнішні джерела $T(References)$, кількість відвідувань за 3 місяці $T(Visit)$, рейтинг у пошукових системах Яндекс та Google $T(Rate)$, місцезнаходження, наявність відео файлів $T(Video)$.

Для цієї сукупності структурних характеристик можна сформувати персоналізований шаблон найбільш прийняттого Web-документу, який буде використано для підвищення ефективності пошуку.

Оцінки структурних характеристик такого шаблону можуть бути представлені як числовими, так і символічними значеннями. Наприклад, бажаний розмір Web-документу може бути зазначено користувачем як фіксована величина. В разі неможливості визначення числових значень характеристики вона розглядається як лінгвістична змінна, що приймає символічні значення «великий, середній та малий», кожне з котрих є нечіткою підмножиною універсальної множини. Для присвоєння таким нечітким підмножинам числових еквівалентів структурних характеристик Web-документу доцільно використовувати функції належності «узагальнений колокол» та «сигмоїдна функція».

Наприклад, для того, щоб побудувати функції належності для нечітких підмножин лінгвістичної змінної «розмір», необхідно виконати наступні операції: вибрати величину для оцінки розміру Web-документу; виконати нормалізацію отриманих значень розмірності Web-документів; визначити границі нечітких підмножин універсальної множини $T(Size)$ (для оцінки діапазонів коливання розмірів Web-документів доцільно вибрати: $big = [1.5 \text{ мб}, \infty)$; $middle = [1.3 \text{ мб}, 2.7 \text{ мб}]$; $small = [0 \text{ мб}, 1 \text{ мб}]$).

Після цього нечіткі підмножини універсальної множини $T(Size)$ визначаються функцією належності наступного вигляду:

$$\mu'' big''(x) = \begin{cases} 1, x \geq 28.5 \\ \frac{1}{1 + e^{-(x-22)}}, x \in [5, 28.5] \end{cases} \quad (6)$$

$$\mu'' middle''(x) = \begin{cases} 0, x \in [5, 20] \\ \frac{1}{1 + (\frac{x-12}{3})^{2*3}}, x \in [20, 27] \end{cases} \quad (7)$$

$$\mu'' small''(x) = \begin{cases} 0, x \in [5, 10] \\ \frac{1}{1 + e^{2(x-5)}}, x \in [10, 7.5] \end{cases} \quad (8)$$

Після формування шаблону структурних характеристик Web-документу його можна використовувати для пошуку необхідної інформації. Процедура пошуку передбачає порівняння цього шаблону зі значеннями структурних характеристик Web-документів, що аналізуються, та подальше створення внутрішнього персоналізованого рейтингу Web-документів.

В другому розділі вперше запропоновано метод інтелектуального пошуку та аналізу інформації в мережі Інтернет, що базується на використанні нових критеріїв, які дозволяють враховувати значущість Web-ресурсів для кожного користувача залежно від його соціального профілю та оцінювати рівень їх зашумленості. Запропонований метод дозволяє обчислювати соціальний індекс (SI) Web-документу для кожного користувача системи пошуку інформації на основі соціального індексування. Для цього використовується значення співвідношення загальної кількості відвідувань до кількості відвідувань для одного досліджуваного типу інтересів користувача CU з діапазоном значень $[0;1]$. При визначенні соціального індексу береться також до уваги середня оцінка привабливості Web-документу MU з діапазоном значень $[0;10]$ за умови для користувачів зі спільним інтересом. Формально соціальний індекс i -го Web-документу для одного досліджуваного типу інтересів користувача визначається наступним чином:

$$SI_i = \sqrt{(MU_t - MU_i)^2 + (CU_t - CU_i)^2}, \quad (8)$$

де MU_t – оцінка значущості шаблону; MU_i – середня оцінка значущості i -го Web-документу для користувачів зі спільним інтересом; CU_t – співвідношення загальної кількості відвідувань шаблону до кількості відвідувань користувачів зі спільним інтересом; CU_i – співвідношення загальної кількості відвідувань i -го Web-документу до кількості відвідувань користувачів зі спільним інтересом.

Використання критерію пошуку інформації за соціальним індексом дозволяє оцінювати значущість Web-документів в залежності від актуальних інтересів кожного користувача.

В **третьому розділі** вдосконалено метод структуризації інформації на основі онтологічного підходу та технології Semantic Web.

Також в цьому розділі пропонується схема побудови системи пошуку релевантних Web-документів та формування адаптивних Web-сторінок, яка використовує розглянуті вище методи. Показано, що для реалізації такої схеми є доцільним створення експертної системи на основі універсального програмного об'єкту, що здійснює пошук та інтелектуальний аналіз даних мережі Інтернет з використанням запропонованих методів та критеріїв.

Розробка правил логічного висновку для бази знань експертної пошукової системи, що розглядається, базується на обробці та структуризації Web-інформації. На рис. 1 наведено запропонований варіант організації бази знань тематичних Інтернет-ресурсів. Ця база знань містить дані, здобуті за допомогою аналізу Інтернет-ресурсів та семантичного опису, створеного згідно зі стандартом Dublin-core. При цьому джерела інформації надходять з плагіну та тематичних каталогів у вигляді посилань на Web-ресурси, які потім обробляються з урахуванням особливостей запропонованої структури бази знань. Наповнення бази знань здійснюється за допомогою онтології, що зберігається у вигляді спеціального файлу з розширенням owl, близьким за своєю структурою до XML-файлу. Для створення онтології використано мову Ontology Web Language (OWL).

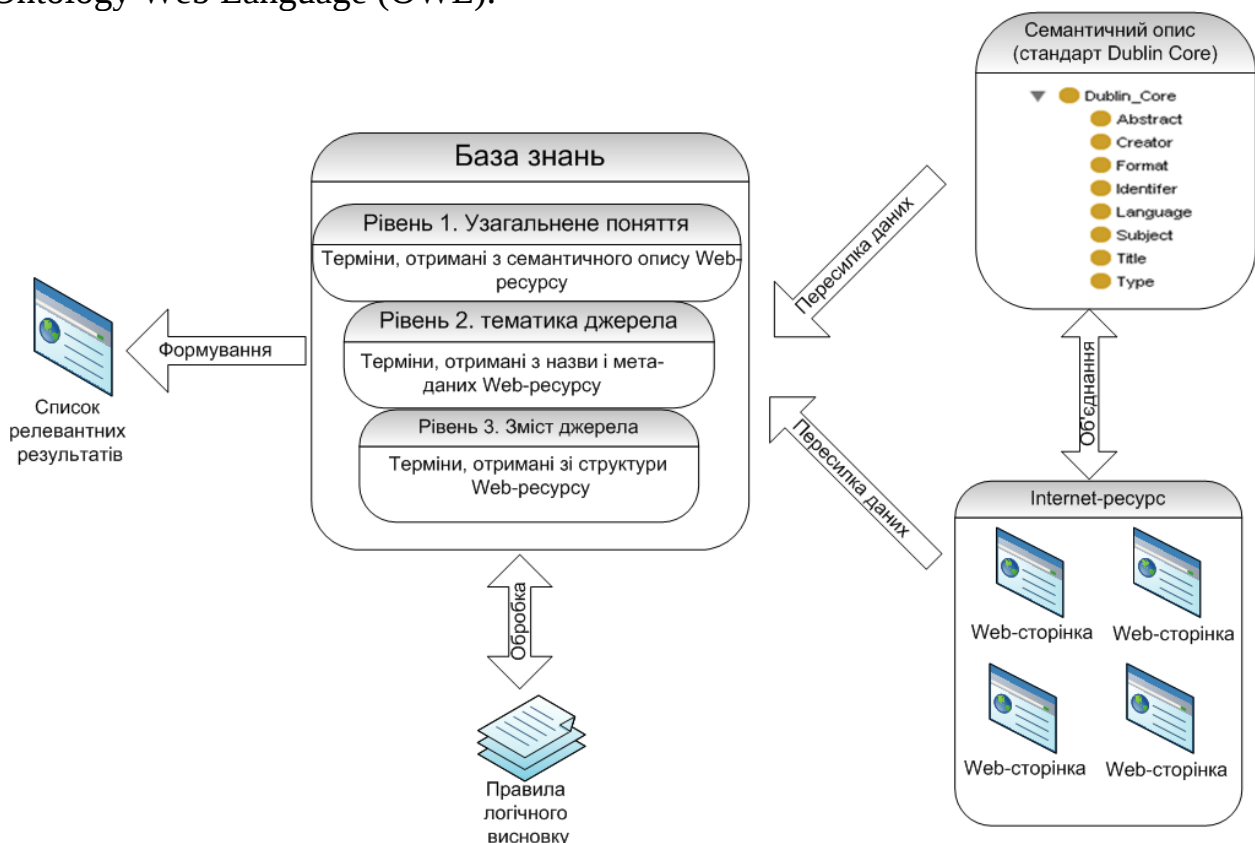


Рисунок 1 – Організація бази знань тематичних Інтернет-ресурсів

Згідно зі своїм призначенням експертна пошукова система, що розглядається, має формувати релевантні результати відповідно до поточних

Підсумковий результат формується експертною системою у вигляді сго рейтингу (PR_i) Web-документів, що залежить від значень соціального індексу і загальної оцінки значущості та визначається за наступною формулою:

$$PR_i = \sqrt{\omega_1(SI_i^2) + \omega_2(d(Optimal)^2_i)}, \quad (9)$$

де i – номер аналізованого Web-документу, $i = \overline{1, m}$ (m – загальна кількість досліджуваних Web-документів); SI_i – значення соціального індексу i -го Web-документу; $d(Optimal)_i$ – загальна оцінка значущості i -го Web-документу; $\omega_{1,2}$ – вагові коефіцієнти.

Розглянута експертна система дозволяє реалізувати розвинений в роботі метод формування адаптивних Web-сторінок шляхом використання універсальної програми-агента, яка забезпечує оперативний доступ користувача до цікавої для нього інформації в мережі Інтернет.

Цей метод передбачає послідовну реалізацію наступних операцій: пошук необхідного джерела інформації для формування персоналізованої динамічної Web-сторінки; обробка та виділення релевантних даних з Web-документу з подальшою перевіркою наявності відповідного електронного документу в сховищі значущої інформації (в разі його відсутності здійснюється обробка Web-сторінки та запис нової інформації до сховища експертної системи); отримана на черговому кроці роботи системи значуща інформація інтегрується до персоналізованої адаптивної Web-сторінки (якщо вона ще відсутня, то спочатку задаються її структура та адреса згідно з запропонованим в дисертації шаблоном, після чого стає можливим поповнення сторінки у звичайному режимі). Процес інтеграції контенту Web-ресурсу, що має актуальний інтерес для користувача, до створеної адаптивної персоналізованої сторінки схематично наведено на рис. 3.

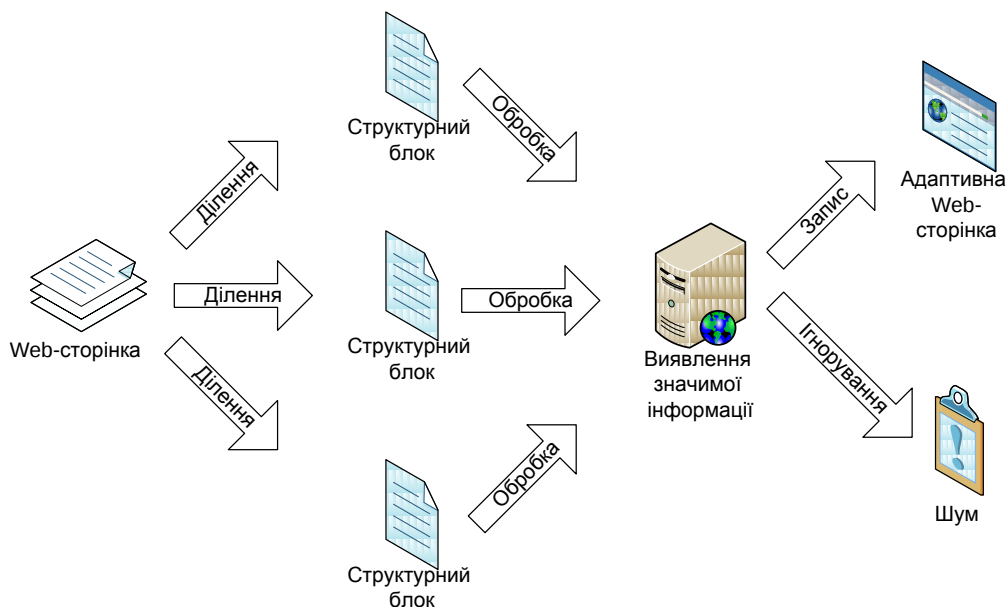


Рисунок 3 – Схема поповнення адаптивної Web-сторінки релевантними даними із сховища бази знань експертної системи

У заключній частині розділу наведено приклади моделювання структури адаптивних персоналізованих Web-сторінок та рекомендації з їх поточної корекції (зокрема, пропонуються порядок та формат генерації Web-даних за результатами обробки сторінок Web-ресурсів зі сховища значущої інформації бази знань експертної системи, що можуть зацікавити користувача).

У **четвертому розділі** розглядаються результати практичної реалізації експертної системи пошуку даних з мережі Інтернет та синтезу адаптивних Web-сторінок (динамічного Web-ресурсу Social Search), що використовує методи, запропоновані в теоретичній частині дисертаційної роботи. Базову модель розробленої системи реалізовано як Web-ресурс з використанням технологій Java Servlets и Java Server Pages, сервера Apache Tomcat 7.0, концепції Model-view-controller (MVC) та шаблону проектування Web-додатків. При цьому релевантним кінцевим результатом роботи системи є адаптивні Web- сторінки, що ранжуються згідно зі значеннями соціального індексу для кожної з досліджуваних тематик.

Система містить наступні компоненти: базу знань, що спроектовано за допомогою спеціалізованого редактора Protégé (версія 3.4.8) згідно з запропонованою у дисертації схемою (передбачено можливість корегування онтології та модифікації базової версії експертної системи з оглядом на актуальні потреби користувача); систему автоматизованого заповнення бази знань; програмний додаток, призначений для тестування працездатності спроектованого динамічного Web-ресурсу Social Search (додаток розроблено за допомогою спеціалізованих інструментів, плагінів та бібліотек, що використовуються у складі фреймворку баз знань Protégé-Owl для підключення до програмного модулю OwlInsert експертної системи).

Розроблену експертну систему було використано для реалізації ряду практичних впроваджень, зокрема, при створенні адаптивних Web-сторінок для газети «Пожежний вісник» видавництва «Пожежна безпека», що здійснюють автоматизовану інтелектуальну обробку та ранжування оперативних даних мережі Інтернет, присвячених тематиці надзвичайних ситуацій, з подальшим формуванням відповідних інформаційних блоків.

Для оцінки якісних характеристик розробленої системи синтезу адаптивних Web- сторінок на основі Web-ресурсу Social Search було здійснено тестування процедур пошуку релевантної інформації в мережі Інтернет для різних типів запитів. Крім розробленого пошукового агенту Social Search в тестовому експерименті досліджувалися також відомі пошукові системи Google, Infostream та експертні системи Similarity/closeness-based resource browser, Personal Searcher. Результати тестування свідчать, що запропонована в дисертації система дещо поступається за швидкістю пошуку системі Google, але має безперечні переваги в порівнянні з нею у якості отриманих результатів за рахунок урахування семантики запитів та персоналізації релевантного результату, орієнтованого на соціальну значущість для користувачів. В той же час, середня швидкість пошуку системи на базі Social Search (3с) є нижчою, ніж відповідний показник спеціалізованих систем Similarity/closeness-based resource

browser, Personal Searcher (відповідно, 5с та 15с). Крім того, запропонована система має найбільші функціональні можливості серед усіх тестованих систем щодо здійснення фільтрації підсумкового результату з погляду запропонованих в дисертаційній роботі критеріїв.

У **додатках** наведено акти про впровадження та практичне застосування результатів, отриманих у дисертаційній роботі.

ВИСНОВКИ

У дисертаційній роботі наведено результати, які відповідно до поставленої мети є вирішенням науково-технічної задачі розробки методів синтезу адаптивних Web-сторінок для підвищення ефективності роботи експертної системи семантичного пошуку та аналізу даних у мережі Інтернет на основі використання нових критеріїв оцінювання їх релевантності.

На підставі проведених теоретичних та експериментальних досліджень отримано такі найсуттєвіші результати:

1. Обґрунтовано доцільність розробки персоналізованої пошукової системи з можливістю генерування адаптивних Web-сторінок з урахуванням актуальних інтересів користувачів.

2. Вперше запропоновано метод визначення структурних характеристик Web-документів на основі використання кластерного аналізу та нечіткої логіки, який дозволяє кількісно оцінювати ступінь привабливості для користувачів даних, що аналізуються, та створювати персоналізований рейтинг електронних джерел інформації.

3. Вперше запропоновано метод інтелектуального пошуку та аналізу інформації в мережі Інтернет, що базується на використанні нових критеріїв, які дозволяють враховувати значущість Web-ресурсів для кожного користувача залежно від його соціального профілю та оцінювати рівень їх зашумленості.

4. Вдосконалено метод структуризації інформації на основі онтологічного підходу та технології Semantic Web. На відміну від існуючих методів він дозволяє створювати та обробляти семантичні описи Web-ресурсів для підвищення ефективності системи семантичного пошуку та аналізу даних.

5. Запропоновано архітектуру експертної пошукової системи, яка не лише дозволяє здійснювати пошук необхідних Web-документів з використанням розроблених у дисертації методів та критеріїв, але й дає змогу користувачеві заздалегідь оцінювати значущість Web-документу без перегляду його тексту.

6. Запропоновано схему формування бази знань експертної пошукової системи з використанням сучасних технологій (OWL, Dublin Core), що сприяє швидкому виводу релевантних результатів пошуку у вигляді списку Web-документів згідно з запитом користувача та з урахуванням його інтересів в рамках заданої предметної області.

7. Отримав подальший розвиток метод формування адаптивних Web-сторінок шляхом використання універсальної програми-агента, яка забезпечує оперативний доступ користувача до цікавої для нього інформації в мережі Інтернет. Його відмінною особливістю є можливість накопичування тематичної інформації згідно з поточним запитом користувача і його

соціальним профілем та подальше інтегрування отриманої значущої інформації з мережі Інтернет до адаптивної Web-сторінки.

8. Розроблені методи було програмно реалізовано та використано для ряду практичних впроваджень, зокрема, для синтезу експертно-пошукової системи газети «Пожежний вісник» та у навчальному процесі Харківського національного університету радіоелектроніки. Результати тестування підтверджують слушність теоретичних висновків дисертаційної роботи.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Почанский О.М. Модель построения адаптивных Web-страниц на основе интеллектуального анализа сети Internet [Текст] / О.М. Почанский // Восточно-Европейский журнал передовых технологий. – 2010. – Вып. 4/7(46). – С. 66-69.

2. Почанский О.М. Извлечение частично-структурированной (значимой) информации из динамических Web-документов [Текст] / О.М. Почанский // Бионика интеллекта. – 2011. – № 3(77). – С. 143-149.

3. Почанский О.М. Применение структурных характеристик Web-документов при оценивании их привлекательности для конечного пользователя [Текст] / О.М. Почанский // Системы обработки информации. – 2011. – Вып. 4 (94). – С. 118-125.

4. Почанский О.М. Социальное индексирование Web-документов для семантического поиска [Текст] / О.М. Почанский // Искусственный интеллект. – 2012. – №1. – С. 112-122.

5. Почанский О.М. Экспертная система семантического поиска релевантных данных и формирования адаптивных Web-страниц [Текст] / О.М. Почанский // Автоматизированные системы управления и приборы автоматики. – 2012. – Вып. 159. – С. 57-66.

6. Почанский О.М. Использование семантического описания для кластеризации блогов и форумов [Текст] / О.М. Почанский // Вопросы прикладной математики и математического моделирования. – 2009. – №1. – С. 321-328.

7. Почанский О.М. Структуризация блогов и форумов на основе их семантического описания [Текст] / О.М. Почанский // Материалы I научно-практической конференции «Интеллектуальные системы. Бионика интеллекта» – Харьков: ХНУРЕ, 2008. – С. 38.

8. Почанский О.М. Использование семантического описания для кластеризации блогов и форумов [Текст] / О.М. Почанский // Материалы V международной научно-практической конференции «Математическое и программное обеспечение интеллектуальных систем» – Днепропетровск: ДНУ, 2008. – С. 320.

9. Почанский О.М. Составление тематического дерева элементов Web 2.0 [Текст] / О.М. Почанский // Материалы 13 международного молодежного форума: «Радиоэлектроника и молодежь в 21 веке» – Харьков: ХНУРЕ, 2009. – С. 121.

10. Почанский О.М. Модель проектирования интеллектуальной системы анализа ресурсов сети Internet [Текст] / О.М. Почанский // Матеріали першої науково-технічної конференції: «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління» –Харків-Київ: ДП «ХНДІ ТМ», ДП «ЦНДІ НіУ», 2010. – С. 67.

11. Почанский О.М. Критерий «чистоты» как основная характеристика модели улучшения работы поискового агента [Текст] / О.М. Почанский // Материалы международной конференции «Математическое и программное обеспечение интеллектуальных систем (MPZIS-2010)» – Днепропетровск, 2010. – С. 180-181.

12. Почанский О.М. Организация базы знаний тематических Internet-ресурсов по средствам их интеллектуального анализа [Текст] / О.М. Почанский // Материалы международной научной конференции «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта» – Евпатория, 2011. – С. 103.

13. Почанский О.М. Разработка персонализированного рейтинга Web-страниц, посвященных одной и той же тематике [Текст] / О.М. Почанский // Материалы первой научно-технической конференции «Вычислительный интеллект (результаты, проблемы, перспективы)» – Черкассы, 2011. – С.363.

14. Почанский О.М. Интеграция интеллектуальной системы анализа данных с социальным профилем пользователя [Текст] / О.М. Почанский // Материалы VI международной научно-практической конференции: «Наука и социальные проблемы общества: информатизация и информационные технологии» – Харьков, 2011. – С. 132-133.

15. Почанский О.М. Мультиагентный подход при генерировании данных из различных источников информации [Текст] / О.М. Почанский // Материалы международной научно-практической конференции: «Информационные технологии и информационная безопасность в науке, технике и образовании». – Севастополь: СЕВНТУ, 2011. – С.155.

16. Почанский О.М. Интеллектуальный анализ Web-данных средствами кластерного анализа [Текст] / О.М. Почанский // Материалы 16 международного молодежного форума «Радиоэлектроника и молодежь в 21 веке». – Харьков: ХНУРЕ, 2012. – С. 41-42.

АНОТАЦІЯ

Почанський О.М. Методи синтезу адаптивних Web-сторінок на основі інтелектуального аналізу інформаційних ресурсів мережі Інтернет. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2013.

Дисертацію присвячено розробці методів синтезу адаптивних Web-сторінок для підвищення ефективності роботи експертної системи семантичного пошуку та аналізу даних у мережі Інтернет на основі

використання нових критеріїв оцінювання їх релевантності. Обґрунтовано доцільність створення персоналізованої пошукової системи з можливістю генерування адаптивних Web-сторінок з урахуванням актуальних інтересів користувачів. Розроблено метод формування та оцінювання структурних характеристик Web-документів, оснований на використанні кластерного аналізу та нечіткої логіки, що дозволяє створювати персоналізований рейтинг електронних джерел інформації. Розроблено метод семантичного пошуку та аналізу інформації, оснований на використанні соціальних індексів, що кількісно характеризують рівень значущості Web-документів для користувачів з різними інтересами. Запропоновано архітектуру та схему формування бази знань експертної пошукової системи з використанням сучасних технологій (OWL, Dublin Core), що сприяє швидкому виводу релевантних результатів пошуку у вигляді списку Web-документів згідно з запитом користувача та з урахуванням його інтересів в рамках заданої предметної області. Розроблені методи було програмно реалізовано та використано для ряду практичних впроваджень.

Ключові слова: адаптивна Web-сторінка, семантичний пошук, соціальний індекс, експертна пошукова система, персоналізований рейтинг, Web-ресурси, інтелектуальний аналіз.

АННОТАЦИЯ

Почанский О.М. Методы синтеза адаптивных Web-страниц на основе интеллектуального анализа информационных ресурсов сети Интернет. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2013.

Диссертация посвящена разработке методов синтеза адаптивных Web-страниц для повышения эффективности работы экспертной системы семантического поиска и анализа данных в сети Интернет на основе применения новых критериев оценивания их релевантности.

Объект исследования – семантический поиск и структурирование информации в Web-пространстве на основе анализа социального профиля пользователя.

Предмет исследования – методы синтеза адаптивных Web-страниц для повышения эффективности работы экспертной системы семантического поиска и анализа данных в сети Интернет на основе применения новых критериев оценивания их релевантности и стандартов Semantic Web.

Для решения сформулированных в работе задач использовались следующие методы: методы семантического поиска информации – для разработки алгоритмов поиска и анализа Web-документов с применением предложенных в диссертации критериев; методы кластерного анализа и нечеткой логики – для формирования и количественного оценивания структурных характеристик Web-документов; методы агентных технологий –

для формирования адаптивных Web-страниц с применением универсальной программы-агента; методы формирования онтологических моделей – для семантического описания Web-ресурсов; методы проектирования сложных Web-приложений – для программной реализации экспертной системы семантического поиска и анализа данных в сети Интернет.

Проведен анализ существующих методов построения информационно-поисковых систем и адаптивных Web-страниц, позволивший обосновать целесообразность разработки эффективной персонализированной поисковой системы с возможностью генерирования адаптивных Web-страниц с учетом актуальных интересов пользователей.

Разработан метод оценивания значимости анализируемых Web-документов, основанный на разделении Web-страниц на структурные блоки и проверке их на наличие шумов. Разработан метод формирования и оценивания структурных характеристик Web-документов, основанный на применении нечеткой логики, позволяет создавать персонализированный рейтинг электронных источников информации.

Разработан метод семантического поиска информации на основе оценивания социальных индексов, численно характеризующих степень значимости Web-документов для пользователей с различными интересами.

Предложены архитектура и схема формирования базы знаний экспертной поисковой системы, которые основаны на использовании современных технологий Semantic Web (OWL, Dublin Core) и позволяют не только осуществлять поиск необходимых данных с применением разработанных методов критериев, но и предварительно оценивать значимость Web-документа без просмотра его текста. Рассмотрена модель формирования персонализированных адаптивных Web-страниц, основанная на взаимодействии пользователя со специальным плагином экспертной системы. Разработана и программно реализована базовая модель системы синтеза адаптивных Web-страниц на основе технологий Java Servlets и Java Server Pages, сервера Apache Tomcat 7.0, шаблона проектирования Web-приложений и концепции Model-view-controller. Тестирование процедур поиска релевантной информации в сети Интернет для различных типов запросов подтвердило целесообразность использования предложенного оригинального программного агента Social Search для персонализированного семантического поиска значимой информации и формирования адаптивных Web-страниц пользователей с учетом их актуальных интересов.

Результаты практической реализации системы синтеза адаптивных Web-страниц (в частности, при создании адаптивных Web-страниц для редакции газеты «Пожарный вестник») подтверждают эффективность методов, предложенных в теоретической части диссертационной работы.

Ключевые слова: адаптивная Web-страница, семантический поиск, социальный индекс, экспертная поисковая система, персонализированный рейтинг, Web-ресурсы, интеллектуальный анализ.

ABSTRACT

Pochanskiy O.M. Synthesis of adaptive Web-pages based on intelligent analysis of information resources on the Internet. – Manuscript.

The thesis is for the Ph.D degree, specialty 05.13.23 – systems and artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkov, 2013.

The thesis is devoted to developing methods for the synthesis of adaptive Web-pages to improve the effectiveness of semantic retrieval and data analysis on the Internet expert system through the use of a new relevance evaluation criterion. The creating personalized search engine with the ability to generate adaptive Web-pages based on actual user interest was justified. Further the method of forming and evaluating the structural characteristics of Web-documents, based on cluster analysis and fuzzy logic, which allows creating a personalized ranking of electronic information sources was designed. Also the method of semantic search and analysis, based on the use of social indexes that quantifies the level of Web-documents significance for users with different interests was developed. And the architecture and scheme of forming the knowledge base of the expert search system using modern technologies (OWL, Dublin Core), which facilitates the rapid output of relevant search results as a list of Web-documents according to user requests and considering its interests within a given subject area were suggested.

All the methods had been implemented as software and applied for several practical implementations.

Keywords: adaptive Web-page, semantic search, social index, search engine expert system, personalized ranking, Web-resources, mining.

Підп. до друку . .13. Формат 60x84/16. Спосіб друку – ризографія

Умов. друк. арк. 1,2. Ціна договірна. Тираж 100 прим.

Зам. .

ХНУРЕ, 61166, Харків, просп. Леніна, 14

Віддруковано в навчально-науковому
видавничо-поліграфічному центрі ХНУРЕ
Харків, просп. Леніна, 14