ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

Краковецький Олександр Юрійович

УДК 681.5:519.876.2

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПТИМІЗАЦІЇ ПОШУКУ ДОКУМЕНТІВ У ВЕБ-СИСТЕМАХ

Спеціальність 05.13.06 – Інформаційні технології

Автореферат

дисертації на здобуття наукового ступеня

кандидата технічних наук

Вінниця – 2011

Дисертацією є рукопис.

Роботу виконано у Вінницькому національному технічному університеті Міністерства освіти і науки, молоді та спорту України.

Науковий керівник:

доктор технічних наук, професор
Дубовой Володимир Михайлович,
Вінницький національний технічний університет,
завідувач кафедри комп'ютерних систем управління

Офіційні опоненти:

доктор технічних наук, професор
Теленик Сергій Федорович,
Національний технічний університет України «КПІ», завідувач кафедри автоматики та управління в технічних системах

доктор технічних наук, професор
Мокін Віталій Борисович,
Вінницький національний технічний університет,
завідувач кафедри моделювання та моніторингу складних систем

Захист відбудеться «19» березня 2011 р. о 9 ^{30} — годині на засіданні спеціалізованої вченої ради Д 05.052.01 у Вінницькому національному технічному університеті за адресою: 21021, м. Вінниця, вул. Хмельницьке шосе, 95, ГНК, ауд. 210.

З дисертацією можна ознайомитись у бібліотеці Вінницького національного технічного університету за адресою: 21021, м. Вінниця, вул. Хмельницьке шосе, 95.

Автореферат розісланий «_15_» лютого 2011 р.

Вчений секретар
спеціалізованої вченої ради С. М. Захарченко

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Щодня Інтернетом користуються мільйони людей, створюються тисячі нових сайтів, мільйони книг та документів переводять в електронний вигляд. Загубитися в такому великому інформаційному просторі дуже легко. Оскільки вимоги до швидкості пошуку, актуальності інформації з кожним днем зростають, то і збільшуються вимоги до методів та алгоритмів пошуку та подання інформації. Процес пошуку та відображення інформації в Інтернеті має ряд особливостей, головними з яких є величезна кількість веб-ресурсів, необхідність врахування семантичних особливостей інформації, вплив великої кількості факторів при пошуку (наявність посилань, рейтинг веб-сторінок і сайтів в цілому, ім'я домену тощо), необхідність врахування особливостей гіпертекстової розмітки та метаінформації. На сьогоднішній день існує велика кількість методів та алгоритмів інформаційного пошуку, проте постійний розвиток цієї галузі та зростання обсягів даних вимагає постійного вдосконалення існуючих методів та розробку якісно нових підходів.

Зв'язок роботи з науковими програмами планами і темами. Дисертаційна робота виконувалась в рамках пріоритетних напрямків розвитку науки і техніки в Україні відповідно до плану науково-дослідних робіт кафедри комп'ютерних систем управління Вінницького національного технічного університету, держбюджетної науково-дослідної роботи «Розробка теорії та методів оптимальних рішень в умовах комбінованої невизначеності» (номер держ. реєстрації 0105U002431), госпдоговірної роботи «Розробка підсистеми оптимізації інформаційного пошуку для корпоративного сайту ІВП "ІнноВінн"» (номер держ. реєстрації 0110U001839).

Мета і завдання дослідження.

Метою роботи є зменшення часу пошуку документів у веб-системах.

Завдання дослідження:

– проаналізувати існуючі методи, алгоритми та технології інформаційного пошуку та інтелектуального аналізу даних;

– розробити метод оцінювання релевантності інформаційних блоків веб-сторінки з точки зору оптимізації для пошукових систем;

– удосконалити математичну модель для оцінювання важливості інформаційних блоків веб-сторінок, що дозволило б підвищити вірогідність знаходження їх основної змістовної частини, та метод очищення веб-сторінок від інформаційного шуму;

– розробити методи, алгоритми та інформаційну технологію, що їх реалізує, для знаходження оптимальної послідовності перегляду результатів пошуку документів у веб-системах;

– перевірити розроблені теоретичні положення, методи та алгоритми на практиці.

Об'єктом дослідження є процес пошуку документів у веб-системах.

Предметом дослідження є методи, моделі та інформаційні технології пошуку, обробки та відображення інформації в Інтернеті.

Методи дослідження базуються на використанні теорії графів для побудови та аналізу взаємозв'язків між веб-ресурсами, теорії інформації для оцінювання контенту, методів та алгоритмів кластеризації та класифікації даних для розбиття ресурсів на семантичні групи, методів оптимізації для знаходження оптимальних шляхів перегляду інформації, технології Data Mining та Text Mining для обробки та аналізу даних, методів та алгоритмів інформаційного пошуку.

Наукова новизна одержаних результатів. В ході розв'язання поставлених задач отримано нові наукові результати:

– вперше розроблено метод знаходження оптимальних шляхів перегляду документів у веб-системах, який, на відміну від існуючих, враховує не лише контент веб-сторінок, але й гіпертекстову структуру, що дає змогу підвищити швидкість багатокрокового процесу отримання інформації в пошуковій системі;

– розроблено метод оцінювання релевантності інформаційних блоків веб-сторінок з

точки зору оптимізації для пошукових систем, який ґрунтується на використанні правил SEO, що дозволив підвищити достовірність ідентифікації типів інформаційних блоків;

– дістав подальшого розвитку метод очищення веб-сторінок від інформаційного шуму, який ґрунтується на розробленій математичній моделі оцінювання важливості інформаційних блоків веб-сторінок і відрізняється від існуючих тим, що враховує ряд семантичних характеристик, що дало змогу підвищити ймовірність правильної ідентифікації основного контенту веб-сторінок;

– розроблено нову інформаційну технологію відображення оптимальної послідовності перегляду результатів пошуку документів у веб-системах, яка відрізняється від існуючих тим, що використовує розроблені методи пошуку оптимальних шляхів перегляду результатів веб-пошуку, оцінювання релевантності інформаційних блоків веб-сторінок з точки зору оптимізації для пошукових систем, очищення веб-сторінок від інформаційного шуму, що дало змогу підвищити ефективність пошуку та зменшити час перегляду документів у веб-системах.

Практичне значення одержаних результатів. На основі запропонованих моделей та методів розроблено алгоритми та програмне забезпечення інформаційної технології оптимізації пошуку документів у веб-системах, а саме:

– алгоритми очищення веб-сторінок від інформаційного шуму, побудови оптимальної послідовності перегляду результатів пошуку у веб-системах, побудови графової моделі взаємозв'язків між сайтами;

– програмне забезпечення для відображення оптимальної послідовності перегляду результатів пошуку у веб-системах, програмні пакети Data Extracting SDK, SmartBrowser, Block Importance Analysis Tool.

Результати дисертаційних досліджень впроваджені:

–на інноваційному виробничому підприємстві «ІнноВінн» у функціональній підсистемі пошуку;

–в компанії TDC LLC в онлайн-системі управління проектами (http://taskpoint.com/);

–в навчальний процес кафедри комп'ютерних систем управління Вінницького національного технічного університету при викладанні дисципліни «Інформаційні технології в системах управління» та «Інтелектуальні технології оптимізації».

Особистий внесок здобувача. Всі результати, які складають основний зміст дисертації, отримані здобувачем самостійно. В роботах, опублікованих у співавторстві, здобувачу належать такі ідеї і розробки: метод пошуку асоціативних правил на основі сильних наборів даних [1], метод прогнозування фінансових часових рядів на основі асоціативних правил [2], метод побудови оптимальних шляхів перегляду результатів веб-пошуку на основі евристичних алгоритмів [4], метод оцінки подібності веб-сторінок [5], математична модель оцінювання важливості інформаційних блоків сайтів [6], підхід щодо інтеграції семантичних даних на веб-сторінки [8], метод SeoRank для оцінки релевантності інформаційних блоків сайтів [9], метод очищення веб-сторінок від інформаційного шуму [10], метод аналізу фінансових даних за допомогою технології Data Mining [11].

Апробація результатів дисертації. Результати дисертаційної роботи доповідались та обговорювались на одинадцяти науково-технічних конференціях: VI міжнародна науково-практична конференція «Інтернет-Освіта-Наука (ІОН-2008)» (м. Вінниця, 2008 р.); IV міжнародна конференція з оптоелектронних інформаційних технологій «Photonics-ODS 2008» (м. Вінниця, 2008); XIII міжнародна конференція з автоматичного управління (Автоматика-2006) (м. Вінниця, 2006); IX міжнародна конференція «Контроль і управління в складних системах (КУСС-2008)» (м. Вінниця, 2008); міжнародна науково-технічна конференція «Давачі, пристрої, системи 2008» (м. Ялта, 2008); друга міжнародна наукова-практична конференція «Інтегровані інтелектуальні робото-технічні комплекси» ІТРК-2009 (м. Київ, 2009); науково-практичних конференціях професорсько-викладацького складу, співробітників і студентів Вінницького національного технічного університету (м. Вінниця, 2006-2010); технічна конференція «Академічні дні Microsoft» (м. Ялта, 2009-2010); технічна

конференція «IT Jam» (м. Київ, 2009, м. Харків, 2010); технічна конференція «Microsoft Live JumpStarts» (м. Варшава, Польща, 2009).

Публікації. Результати теоретичних і експериментальних досліджень викладені в 12 наукових працях, серед яких десять статей в фахових журналах, що входять до переліку ВАК України, одні тези, свідоцтво про реєстрацію авторського права на комп'ютерну програму.

Структура та обсяг дисертації. Дисертаційна робота складається із вступу, 4 розділів, висновків, списку використаних джерел (158 найменувань) і додатків. Основний зміст викладено на 116 сторінках друкованого тексту, містить 38 рисунки, 15 таблиць. Загальний обсяг дисертації 150 сторінки. Додатки містять окремі лістинги програм та акти впровадження результатів роботи і викладені на 39 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність розроблення інформаційної технології оптимізації пошуку документів у веб-системах. Сформульовано мету і задачі досліджень, наукову новизну і практичне значення отриманих результатів. Подано відомості про особистий внесок автора, апробацію результатів роботи та публікації.

У **першому** розділі проведено дослідження стану задачі пошуку та відображення інформації у веб-системах. Проведено аналіз проблеми підвищення ефективності пошуку документів в Інтернеті, що показав наявність великої кількості проблем, з якими зіштовхуються Інтернет-користувачі в процесі інформаційного пошуку, а саме: велика кількість інформаційного шуму та дублікатів, відсутність кластеризації результатів пошуку, великий час пошуку тощо. Проаналізовано методи пошуку оптимального маршруту перегляду результатів веб-пошуку, а саме точні методи, генетичні алгоритми та мурашині алгоритми. Порівняння часу для вирішення таких задач різними методами показало, що мурашині алгоритми працюють найшвидше. Крім того, поведінка мурах інтуїтивно схожа на поведінку користувачів при інформаційному пошуку, що робить мурашині алгоритми оптимальним інструментом для розв'язання поставленої задачі.

Зміст веб-сторінок є важливою характеристикою для аналізу та побудови оптимальних маршрутів перегляду результатів пошуку документів у веб-системах, оскільки вони не повинні містити веб-сторінки, контент яких дублюється на інших сторінках, кількість інформаційного шуму має бути мінімальною, а основний контент – релевантним предмету пошуку. Тому оцінка веб-сторінок на предмет дублювання інформації та її новизни є необхідним етапом при побудові оптимальних маршрутів перегляду результатів інформаційного пошуку. При аналізі методів та алгоритмів оцінювання змісту веб-сторінок було розглянуто методи оцінювання кількості інформації, подібності текстів для визначення дублікатів та їх новизни. Для розв'язання задачі знаходження дублікатів найкращим чином підійшов метод шинглів, основна ідея якого полягає у розбитті текстів**,** що порівнюються**,** на вибрані з тексту послідовності слів (шингли), для кожного з яких розраховується контрольна сума.

Міра близькості двох текстових документів $sim(D_i, D_j)$ визначалась на основі апарату умовних ймовірностей, а саме, як добуток ймовірності того, що випадкове слово *w* входить в документ D_i за умови, що воно входить в документ D_j, помножене на ймовірність входження цього слова в документ D_j

$$sim(D_i, D_j) = P(w \in D_i \mid w \in D_j) P(w \in D_j). \quad (1)$$

Тоді параметр новизни New_i документа D_i можна записати так:

$$New_i = \frac{Rank_i \cdot sim(D_i, PlusDic)}{\log(i+1) \sum_{j=1}^{N} sim(D_i, D_j)}, \quad (2)$$

де N – загальна кількість веб-документів; D_1 – поточний документ; D_i – i-й документ; $PlusDic$ – словник; $sim(D_i, D_j)$ – міра близькості документів i і j; $sim(D_i, PlusDic)$ – міра близькості документу i і словника; $Rank_i$ – ранг i-го документа.

Кластеризація результатів пошуку є важливим етапом для підвищення ефективності інформаційного пошуку. На основі проведеного аналізу методів кластеризації було вирішено використовувати нечіткий метод c-means. Для застосування класичних методів кластеризації на графах було вирішено завдання побудови семантичного графу для набору веб-документів.

Аналіз літературних джерел із вказаної проблематики показує перспективність досліджень завдань інформаційного пошуку. В результаті проведеного аналізу сформульовані такі завдання: проаналізувати існуючі методи, алгоритми та технології інформаційного пошуку та інтелектуального аналізу даних; розробити метод оцінювання релевантності інформаційних блоків веб-сторінки з точки зору оптимізації для пошукових систем; удосконалити математичну модель для оцінювання важливості інформаційних блоків веб-сторінок, що дозволив би підвищити вірогідність знаходження їх основної змістовної частини, та метод очищення веб-сторінок від інформаційного шуму; розробити методи, алгоритми та інформаційну технологію, що їх реалізує, для знаходження оптимальної послідовності перегляду результатів пошуку у веб-системах; перевірити розроблені теоретичні положення, методи та алгоритми на практиці.

У другому розділі розроблено теоретичні основи аналізу веб-сторінок та взаємозв'язків між ними, а також знаходження оптимальних шляхів перегляду результатів веб-пошуку з використанням методів інтелектуального аналізу, семантичних та синтаксичних методів аналізу текстів, регресійного аналізу та методів оптимізації.

Для оцінювання коефіцієнта унікальності використовується вираз

$$\alpha = \frac{n-m}{n}, \tag{3}$$

де n – загальна кількість речень основного контенту оригінальної веб-сторінки, m – кількість речень, знайдених в інших веб-сторінках з набору.

Для оцінювання шляхів перегляду результатів використовується поняття *інформативності* шляху, яке розраховується за допомогою виразу

$$I_L = \sum_{i \in L} (New_i) \cdot \alpha_{i/1,..i-1} \frac{I_i}{I_j} r_i, \quad L \subset P, \tag{4}$$

де L – шлях перегляду – впорядкована підмножина результатів пошуку; I_i – кількість інформації змістовної частини веб-сторінки; I_j – загальна кількість інформації веб-сторінки; α_i – коефіцієнт унікальності веб-сторінки в межах шляху. Одна і та ж веб-сторінка може мати різні значення коефіцієнта унікальності в залежності від порядку розміщення та наявності інших веб-сторінок, що входять в цей шлях (таким чином шляхи, що будуть складатися з одних і тих самих веб-сторінок, але які будуть розміщені в різних порядках, матимуть різні значення інформативності); r_i – коефіцієнт релевантності веб-сторінки щодо пошукового запиту. Його значення розраховується за допомогою метрики TF-IDF; New_i – необов'язковий параметр, що характеризує новизну контенту, може використовуватися для оцінки новин та інформації, яка швидко втрачає свою актуальність.

Задача пошуку оптимальних шляхів перегляду результатів веб-пошуку звучить таким чином: необхідно знайти такий шлях, в якого відношення інформативності шляху до часу його перегляду буде максимальним. Формально задача описується наступним виразом:

$$f \to \max\left(\frac{I_L}{t}\right), \qquad (5)$$

де I_L – інформативність шляху (4); t – час, що необхідний для перегляду шляху, що розраховується як відношення загальної кількості інформації шляху I_{All} до середньої швидкості перегляду інформації v, тобто $t = \frac{I_{All}}{v}$.

Для визначення веб-сторінок з дублюванням основного контенту, було вирішено задачу очищення веб-сторінок від інформаційного шуму. Для цього веб-сторінки розбивалися на інформаційні блоки – найпростіші інформаційні одиниці з точки зору змістовної подачі матеріалів. Це завдання було вирішено за допомогою методу VIPS від компанії Microsoft. Наступним етапом стало оцінювання інформаційних блоків та знаходження основного контенту (рис. 1), що було досягнуто за рахунок класифікації інформаційних блоків за їх важливістю (табл. 1).

Таблиця 1

Типи інформаційних блоків за їх важливістю

	Наповнення блока	**Ступінь важливості**	**Числове значення**
1.	Рекламні блоки, інформація про сайт, анонси, рекомендації	Не важливий	5
2.	Блоки з навігацією, меню, «шапки» сайтів, форми для введення інформації, пошуку, список суміжних тем	Мало важливий	15
3.	Основне наповнення	Важливий	100

Як навчальні дані для дослідження моделі оцінювання важливості інформаційних блоків був використаний корпус веб-документів, наданий компанією Reuters (http://trec.nist.gov/data/reuters/reuters.html). Цей корпус текстів містить 806791 англомовних текстів новин, що були зібрані протягом 1996-1997 рр. в форматі NewsML.

Для підвищення достовірності оцінки було запропоновано метод *SeoRank* для визначення релевантності інформаційних блоків веб-сторінки щодо її основного змісту, який представлений на веб-сторінці у вигляді інформації у метатегах. На відміну від існуючих методів оцінки релевантності (наприклад, PageRank), SeoRank не розглядає релевантність інформаційних блоків щодо конкретних пошукових запитів і не враховує зовнішні параметри, такі як взаємозв'язки між ресурсами, фізичну доступність ресурсу, відповідність веб-стандартам тощо, а дає можливість оцінити інформаційні блоки в межах конкретної веб-сторінки.

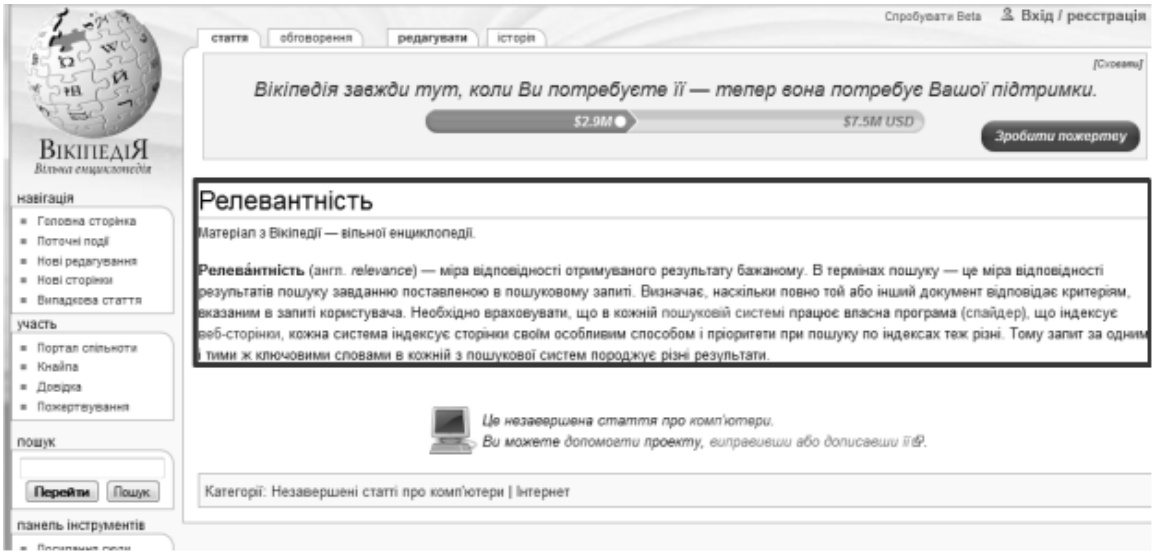

Рис. 1. Приклад веб-сторінки, на якій виділено основний контент

Формально *SeoRank* обчислюється за допомогою адитивного виразу

$$SeoRank = \sum_{i=1}^{4} \alpha_i r_i, \tag{6}$$

де r_i – значення параметра; α_i – вага параметра, причому $\sum_{i=1}^{4} \alpha_i = 1$.

Для обчислення *SeoRank* використовуються такі параметри:

– релевантність заголовку веб-сторінки («title») до тексту інформаційного блока r_1 – відношення кількості входжень слів з заголовку у текст блока до загальної кількості слів блока;

– релевантність ключових слів веб-сторінки («meta keywords») до тексту інформаційного блока r_2 – відношення кількості входжень ключових слів у текст блока до загальної кількості слів блока;

– релевантність слів з опису веб-сторінки («meta description») до тексту інформаційного блока r_3 – відношення кількості входжень слів з опису веб-сторінки у текст блока до загальної кількості слів блока;

– релевантність заголовків веб-сторінки («headers») до тексту інформаційного блока r_4 – відношення кількості входжень слів з заголовків («H1»–«H6») веб-сторінки до загальної кількості слів з заголовків блока.

В результаті дослідження було отримано наступну регресійну модель оцінки важливості інформаційних блоків:

$$\begin{aligned} y = {} & 0{,}324 \cdot x_3 - 0{,}249 \cdot x_5 - 0{,}008 \cdot x_6 + 0{,}056 \cdot x_7 - 0{,}594 \cdot x_{12} - \\ & - 0{,}267 \cdot x_{14} + 0{,}002 \cdot x_{16} + 0{,}305 \cdot x_{17}. \end{aligned} \tag{7}$$

де x_3 – відношення кількості слів блока, що входять у речення, до загальної кількість слів, що входять у речення; x_5 – відношення кількості слів блока, що є посиланнями, до загальної кількості слів, що є посиланнями; x_6 – відношення кількості зображень блока до загальної кількості зображень; x_7 – відношення кількості зображень блока, що є посиланнями, до загальної кількості зображень, що є посиланнями; x_{12} – відношення кількості заголовків (*H1–H6*) до загальної кількості заголовків; x_{14} – відношення кількості слів блока, що є

елементами списків, до загальної кількості слів, що є елементами списків; x_{16} – коефіцієнт читабельності тексту; x_{17} – коефіцієнт релевантності інформаційного блока *SeoRank*.

Для проведення контент-аналізу веб-сайтів та побудови семантичних графів на основі корпусу веб-докуметів в роботі запропонована концепція сильних наборів даних. Асоціативним правилом називається імплікація $X \Rightarrow Y$, де $X \subset I$, $Y \subset I$ и $X \cap Y = \varnothing$. Правило $X \Rightarrow Y$ має підтримку s (support), якщо s% транзакцій з *D* містять $X \cup Y$, supp($X \Rightarrow Y$)=supp($X \cup Y$). Правило $X \Rightarrow Y$ справедливо з достовірністю (confidence) c, якщо c % транзакцій з усієї множини *D*, що містять набір *X*, також містять набір *Y*, тобто conf($X \Rightarrow Y$)=supp($X \cup Y$)/supp(*X*).

Асоціативне правило $X \Rightarrow Y$ називається *допустимим* для заданих σ і τ, якщо виконуються такі умови:

$$\begin{aligned} &Supp(X) \geq \sigma, \\ &Supp(X \cup Y) \geq \sigma, \\ &\frac{Supp(X \cup Y)}{Supp(X)} \geq \tau, \end{aligned} \qquad (8)$$

де σ, τ – відповідно задані мінімальна підтримка і достовірність, причому $0 \leq \sigma, \tau \leq 100$.

Якщо будь-які набори даних X і Y множини A утворюють допустимі асоціативні правила $X \Rightarrow Y$ для заданих σ і τ, то множину A будемо називати *сильним набором даних*.

Формально, нехай маємо множину сильних наборів даних $S = \{s_1, s_2, ..., s_n\}$ і набір асоціативних правил $R = \{s_i \rightarrow s_j\}, s_i, s_j \in S$. Кожний елемент множини $s_i \in S$ є вершиною графу, а асоціативні правила $r_{i,j} \in R$ – ребрами семантичного графу.

Вага кожного ребра графу характеризується відповідними значеннями достовірності та підтримки асоціативного правила. Таким чином, ця методика дозволяє будувати семантичні графи для заданого набору веб-сторінок, що дає змогу звести задачу кластеризації результатів веб-пошуку до задачі кластеризації на графах, методи розв'язання якої були розглянуті в першому розділі.

У третьому розділі розроблено архітектуру інформаційної технології оптимізації пошуку документів у веб-системах. Інформаційна технологія включає в себе математичні моделі, методи та алгоритми пошуку оптимальних рішень, методи контент-аналізу, а також програмне забезпечення, що реалізує ці методи.

Інформаційна технологія складається з декількох функціональних підсистем (ФП), які відповідають за отримання даних з зовнішніх джерел, обробку та аналіз даних, прийняття рішення та візуалізацію рішень. ФП отримання даних відповідає за отримання інформації з пошукових систем, бази даних, сторонніх сервісів, що необхідні для функціонування інформаційної технології.

Система функцій інформаційної технології розроблена у вигляді UML-діаграми варіантів використання, що зображена на рис. 2.

Рис. 2. Діаграма варіантів використання інформаційної технології

На відміну від існуючих, розроблена інформаційна технологія дозволяє не лише відображати лінійний список релевантних веб-ресурсів, але й надавати рекомендації користувачам щодо маршруту їх перегляду. Це дає змогу зменшити час на пошук та перегляд низькорелевантної інформації. В цілому, робота користувача схожа на звичну роботу з пошуковими системами, що робить її зручною у використанні. Перевагою розробленої інформаційної технології є незалежність від джерела даних, що дає можливість використовувати її разом з пошуковими системами, інформаційними ресурсами (такими як Вікіпедія, MSDN тощо) і іншими веб-сайтами, що містять функцію пошуку. Крім того, додатковими можливостями ІТ є можливість відображення графу взаємозв'язків між сайтами, кластеризації результатів пошуку, а також відображення лише основного контенту при перегляді результатів пошуку.

Проведені дослідження моделей і методів прийняття рішень дозволяють узагальнити алгоритм знаходження оптимальних шляхів перегляду результатів веб-пошуку, що складається з таких етапів:

1. Користувач (опційно) обирає налаштування, за яким здійснювати пошук та відображення результатів та вводить пошуковий запит. Результатом цього етапу є список результатів веб-пошуку та пов'язаних веб-сайтів.
2. Веб-сторінки очищаються від інформаційного шуму.
3. З набору результатів видаляються веб-сторінки, які містять контент, що дублюється.
4. Проводиться кластеризація результатів пошуку і якщо пошуковий запит відноситься до різних областей знань – користувачеві пропонується уточнити свій запит.
5. Будується граф взаємозв'язків результатів веб-пошуку.
6. Здійснюється пошук оптимального шляху перегляду.
7. Результати відображаються на екрані.

На другому етапі для підвищення ефективності роботи інформаційної технології для аналізу веб-сторінок використовується лише основний контент, який знаходиться за допомогою такого алгоритму:

1. На вхід подається веб-сторінка, яка ділиться на окремі інформаційні блоки за допомогою методу VIPS. На основі регресійної моделі оцінки важливості інформаційних блоків сайтів для кожного блока розраховуються числові значення важливості.
2. За допомогою нечіткого методу кластеризації c-means (на основі числових значень оцінки важливості) блоки діляться на три кластери (відповідно до трирівневої системи оцінки важливості, див. табл. 1). Необхідність проведення кластеризації пояснюється тим, що числові значення важливості інформаційних блоків, які є основним контентом, можуть бути різними для різних веб-сторінок і тому визначити чіткий діапазон значень, який точно вказував би на тип блока, не можна. Використання методу кластеризації дозволяє визначити типи блоків в рамках окремої веб-сторінки.
3. На виході отримуємо інформаційні блоки, що були визначені як основний контент.

Ефективність запропонованого алгоритму було перевірено на вибірці із 50 тис. веб-документів з набору документів Reuters. Кожну веб-сторінку було розбито на інформаційні блоки, а кожен з блоків було оцінено за допомогою регресійної моделі. Для всіх веб-сторінок проведено кластеризацію за допомогою методу c-means. В результаті дослідження ефективності алгоритму знаходження основного контенту було знайдено, що в 94,36 % алгоритм вірно виділив основний контент. Таким чином, отримані результати свідчать про високу точність роботи запропонованого алгоритму.

Для знаходження оптимальних шляхів перегляду результатів веб-пошуку (шостий етап) використано метод мурашиних колоній. Моделювання поведінки мурах пов’язане з розподілом феромону на шляху – ребрі графу. Ймовірність того, що мураха піде по конкретному ребру пропорційна кількості феромону на цьому ребрі, а кількість феромону, що відкладається на шляху, пропорційна *видимості* між двома веб-ресурсами.

Видимість визначається як евристичне бажання комахи відвідати ресурс j, якщо вона знаходиться на ресурсі i. Значення видимості визначимо як значення коефіцієнта подібності α (3). Чим меншим буде його значення, тобто $V = \frac{1}{\alpha}$ (що означатиме більшу унікальність контенту), тим більше феромону буде відкладено на ребрі, що їх з’єднує, і відповідно, тим більше мурах буде рухатися по цьому маршруту. Для попередження знаходження локально оптимальних рішень необхідно додати зворотній зв'язок у вигляді випаровування феромону. Це дасть змогу отримати стійке оптимальне рішення. Опишемо правила поведінки мурах при виборі маршруту. Мурахи можуть рухатися по тих самих ребрах декілька разів, якщо це необхідно, проте це є небажаним, так як призводить до дублювання фрагментів шляху. Крім того, в спірних моментах мурахи повинні віддавати перевагу тим ребрам, які вони ще не відвідали (якщо такі є). Позначимо через $J_{i,k}$ список ресурсів, які ще не відвідувала комаха k, що знаходиться на i-му ресурсі. «Нюх» комах визначається їх здатністю відчувати запах феромону, що підтверджує бажання відвідати ресурс j із ресурсу i на основі досвіду інших мурах. Кількість феромону на ребрі (i, j) в певний момент t позначимо як $\tau_{i,j}(t)$. Кількість феромону, що відкладе мураха k на ребрі (i, j), перейшовши з ресурсу i на ресурс j, визначається формулою

$$\Delta\tau_{ij,k}(t) = \begin{cases} f_{p_k(t)}, (i, j) \in p_k(t); \\ 0, (i, j) \notin p_k(t), \end{cases} \qquad (9)$$

де $p_k(t)$ – шлях, що пройшла мураха k на момент часу t; $f_{p_k(t)}$ – числова міра оптимальності шляху p.

Правило, що визначає ймовірність переходу *k-oї* мурахи з документа *i* в документ *j* на *t*-й ітерації:

$$\begin{cases} P_{ij,k}(t) = \dfrac{\left[\tau_{ij}(t)\right]^{\alpha} \cdot \left[\eta_{ij}\right]^{\beta}}{\sum_{l \in J_{i,k}} \left[\tau_{il}(t)\right]^{\alpha} \cdot \left[\eta_{il}\right]^{\beta}}, \; if \;\; j \in J_{i,k}; \\ P_{ij,k}(t) = 0, \; if \;\; j \notin J_{i,k}, \end{cases} \tag{10}$$

де α і β – два регульовані параметри, що задають ваги сліду феромону і видимості при виборі маршруту. Правило оновлення феромону має вигляд

$$\tau_{ij}(t+1) = (1-p) \cdot \tau_{ij}(t) + \Delta\tau_{ij}(t), \tag{11}$$

де $\Delta\tau_{ij}(t) = \sum_{k-1}^{m} \Delta\tau_{ij,k}(t)$; m – кількість мурах в колонії.

Алгоритм пошуку шляху є адаптивним до параметрів пошуку, заданих користувачем. Якщо користувач обирає швидкий пошук, то для побудови графу використовується менша кількість результатів пошуку, в іншому випадку кількість рівнів зв'язків сайтів може бути збільшена.

Дослідження ефективності розробленої інформаційної технології було проведено за наступною методикою: десять респондентів повинні були знайти вичерпну інформацію по десяти темам, що відносилися до різних областей знань. Їм дозволялося робити необмежену кількість запитів для знаходження необхідних документів. Результатом їх пошуку був список найбільш релевантних (за їх думкою) веб-документів, які потім порівнювалися з результатами, отриманими за допомогою інформаційної технології. Результати дослідження ефективності розробленої інформаційної технології наведено в табл. 2.

Таблиця 2

Результати дослідження ефективності розробленої ІТ

Назва методу	Відношення важливої / неважливої інформації	Сер. к-сть переглянутих сторінок за один запит	Час пошуку (перегляду), хв.	Значення критерію оптим., МБ/хв.
Класичний пошук	0,3	12	5,7	0,189
Пошук за допомогою ІТ	0,7	6,3	3,7	0,358
Статистичні дані Яндексу	0,3	10	6	0,15

У четвертому розділі надані результати практичної реалізації інформаційної технології.

Для розробки веб-систем найчастіше використовують трирівневу систему, що складається з серверної частини (back-end), бізнес-логіки (business logic) та клієнта (front-end). Саме такий вигляд архітектури був обраний для розробки програмного забезпечення, що реалізує інформаційну технологію.

Останнім часом набув популярності архітектурний шаблон Модель-Вигляд-Контролер (MVC, Model-View-Controller) для розробки програмного забезпечення. Для побудови веб-додатків було обрано технологію ASP.NET MVC, що в повній мірі реалізує цей шаблон. Архітектура програмного забезпечення інформаційної технології має вигляд, що зображений на рис. 3.

Інформаційна система побудована таким чином, щоб підтримувати велику кількість джерел даних (провайдерів). Для цього було розроблено програмний інтерфейс *IDataSource*, що має дві обов'язкові функції – *GetSearchResults(searchQuery)* і *GetRelatedResults(result)*, які необхідно реалізувати для того, щоб система могла працювати з новим постачальником

даних. Функція *GetSearchResults* відповідає за отримання списку з результатами пошуку, в той час як *GetRelatedResults* відповідає за отримання інформації про структуру взаємозв'язків між веб-сторінками – результатами пошуку. В результаті за допомогою цих методів програмне забезпечення будує граф взаємозв'язків для подальшого аналізу.

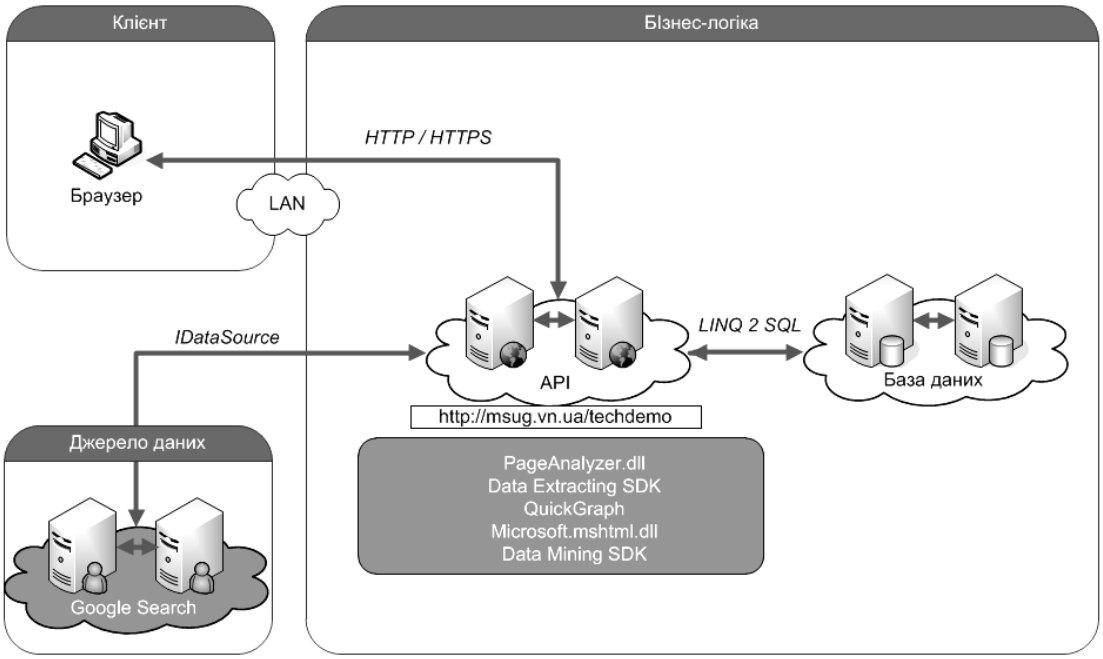

Рис. 3. Архітектура програмного забезпечення IT

Побудова графу взаємозв'язків між веб-сторінками реалізовано за допомогою бібліотеки *QuickGraph*. Зокрема, в роботі використовується структура даних *BidirectionalGraph* – реалізація направленого графу.

Інформаційна технологія оперує даними у форматі HTML, тому однією з прикладних задач було створення бібліотеки для роботи з ними. Для цих цілей було розроблено бібліотеку Data Extracting Software Development Kit (SDK) –набір програмних засобів для роботи з даними, їх обробкою та аналізом.

Клієнт реалізований у вигляді веб-системи, що складається з пошукової форми та результатів пошуку. Остання розділена на дві області – для відображення лінійного списку результатів пошуку та у вигляді графу (рис. 4).

Результати дисертаційних досліджень були впроваджені та отримали практичну реалізацію в компанії ІВП «ІнноВінн», що займається розробкою рішень для телекомунікаційних компаній, а також різноманітного програмного забезпечення. Результати впроваджені у вигляді програмного забезпечення, що включає в себе підсистему оптимізації інформаційного пошуку, що була створена на основі методу пошуку оптимальних шляхів перегляду результатів пошуку та програмних компонентів для обробки і аналізу гіпертекстової інформації.

В процесі впровадження інформаційної технології в онлайн-систему управління проектами компанії TDC LLC. було розроблено додаткові семантичні атрибути, що дало змогу спростити процес індексації. Для цього було запропоновано використовувати метаінформацію на рівні окремих інформаційних блоків. Використання такого підходу при розробці веб-системи дало змогу зменшити час на пошук основного контенту.

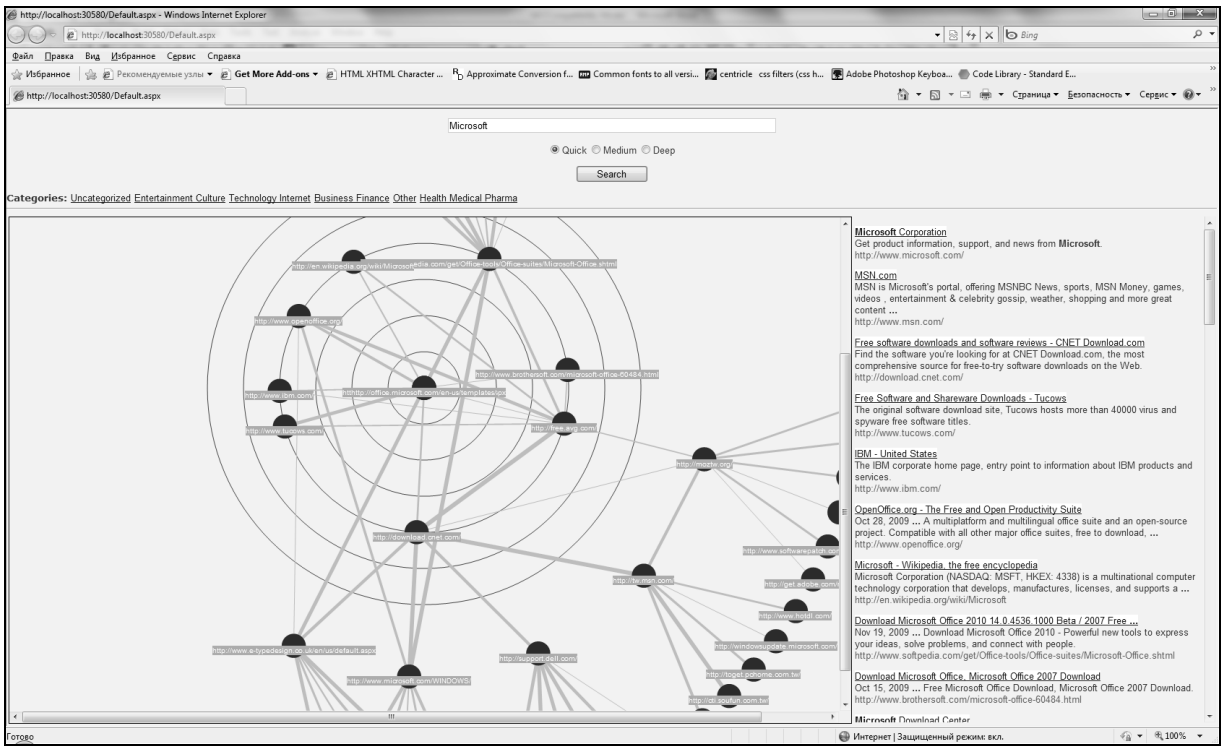

Рис. 4. Зовнішній вигляд системи – реалізації інформаційної технології

Матеріали досліджень дисертаційної роботи були використані у Вінницькому національному технічному університеті під час читання лекцій та проведення практичних занять з курсів «Інформаційні технології в системах управління» та «Інтелектуальні технології оптимізації».

ВИСНОВКИ

У дисертаційній роботі наведено теоретичне узагальнення і нове вирішення актуальної наукової задачі, яка полягає у розробці інформаційної технології оптимізації пошуку документів у веб-системах з метою зменшення часу пошуку документів. В результаті проведеного теоретичного аналізу сучасних поглядів на цю проблему в літературних та Інтернет-джерелах і виконаних досліджень сформульовані та обґрунтовані такі наукові висновки і практичні результати:

1. Виходячи з проведеного аналізу сучасного стану досліджень в області оптимізації інформаційного пошуку було встановлено, що Інтернет-користувачі зіштовхуються з цілою низкою проблем, серед яких велика кількість контенту, що дублюється, відсутність розбиття результатів веб-пошуку за тематиками та велика кількість інформаційного шуму при перегляді веб-сторінок, що значно збільшують час пошуку та перегляду документів. Ці та інші проблеми вказують на те, що розроблення інформаційної технології оптимізації пошуку документів у веб-системах з метою зменшення часу пошуку при багатокроковому перегляді є актуальною теоретичною і прикладною задачею.

2. В роботі запропоновано метод пошуку оптимальних шляхів перегляду документів у веб-системах на основі евристичних алгоритмів, який відрізняється від існуючих тим, що враховує лише основний контент веб-документів і гіпертекстову структуру між ними. Це дало змогу підвищити швидкість багатокрокового процесу отримання інформації в пошуковій системі.

3. На основі правил оптимізації для пошукових систем (SEO) було розроблено метод SeoRank для оцінювання релевантності інформаційних блоків веб-сторінок щодо основного змісту. Метод ґрунтується на припущенні, що метаінформація веб-сторінки відповідає інформаційному блоку, який містить переважно змістовний контент, та використовується в

моделі оцінювання важливості інформаційних блоків веб-сторінок.

4. Дістав подальшого розвитку метод очищення веб-сторінок від інформаційного шуму, який ґрунтується на розробленій математичній моделі оцінювання важливості інформаційних блоків веб-сторінок. Завдяки тому, що метод враховує цілу низку характеристик інформаційних блоків – як синтаксичних (на основі синтаксису мови розмітки HTML), так і семантичних, (зокрема, SeoRank), ймовірність правильної ідентифікації основного контенту веб-сторінок підвищилась.

5. Розроблено нову інформаційну технологію відображення оптимальної послідовності перегляду результатів пошуку у веб-системах, яка відрізняється від існуючих тим, що використовує розроблені методи пошуку оптимальних шляхів перегляду результатів веб-пошуку, оцінювання релевантності інформаційних блоків веб-сторінок з точки зору оптимізації для пошукових систем, очищення веб-сторінок від інформаційного шуму. Це дозволило зменшити час перегляду документів у веб-системах в середньому на 25 %, а кількість отриманої релевантної інформації за одиницю часу – збільшити в 1,8 разів.

6. На основі запропонованих моделей та методів розроблено алгоритми та програмне забезпечення інформаційної технології прийняття рішень системи керування розподіленим об’єктом, а саме алгоритми очищення веб-сторінок від інформаційного шуму, побудови оптимальних шляхів перегляду веб-ресурсів, побудови графової моделі взаємозв’язків між сайтами, програмне забезпечення для пошуку оптимальних шляхів перегляду результатів інформаційного пошуку в веб-системах, програмні пакети Data Extracting SDK, Data Mining SDK, SmartBrowser, Block Importance Analysis Tool.

7. Результати дисертаційних досліджень впроваджені на інноваційному виробничому підприємстві «ІнноВінн» у функціональній підсистемі пошуку; в компанії TDC LLC в онлайн-системі управління проектами (http://taskpoint.com/); в навчальний процес кафедри комп’ютерних систем управління Вінницького національного технічного університету при викладанні дисципліни «Інформаційні технології в системах управління» та «Інтелектуальні технології оптимізації».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Краковецький О.Ю. Метод отримання знань у вигляді асоціативних правил на основі пошуку «сильних» наборів / С.Б. Дубіненко, О.Ю. Краковецький // Вісник Черкаського державного технологічного університету. – Спецвипуск, 2007. – С. 77–79.

2. Краковецький О.Ю. Прогнозування фінансових рядів на основі пошуку асоціативних правил / І.В. Богач, О.Ю. Краковецький, Б.М. Стрихалюк // Збірник наукових праць Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова. – К., 2007. – №35. – С.126–130.

3. Краковецький О.Ю. Метод пошуку асоціативних правил на основі сильних наборів даних і FP-дерева [Електронний ресурс] / О.Ю. Краковецький // Наукові праці ВНТУ. – 2008. – №1. – Режим доступу : http://www.nbuv.gov.ua/e-journals/vntu/2008-1/uk.files/08kauiaf_uk.pdf. – Заголовок з екрану.

4. Краковецький О. Ю. Метод побудови оптимальних шляхів перегляду результатів веб-пошуку на основі евристичних алгоритмів / В. М. Дубовой, О. Ю. Краковецький, О. В. Глонь // Інформаційні технології та комп’ютерна інженерія. – 2008. – №3 (13). – С. 58–61.

5. Краковецький О. Ю. Метод оцінки подібності веб-сторінок / В. М. Дубовой, О. Ю. Краковецький, О. В. Глонь // Оптико-електронні інформаційно-енергетичні технології. – 2008. – №2 (16). – С. 5-9.

6. Краковецький О. Ю. Факторний аналіз оцінки важливості інформаційних блоків сайтів / В. М. Дубовой, О. Ю. Краковецький, О. В. Глонь // Вісник Вінницького політехнічного інституту. – 2008. – №6. – С.103–107.

7. Краковецький О. Ю. Метод кластеризації на основі кластерів, розподілених за нормальним законом / О. Ю. Краковецький // Інформаційні технології та комп’ютерна інженерія. – 2008. – №1(11). – С. 56–60.

8. Krakovetskiy O. Y. The integration of semantic data in hypertext networks / V. M. Dubovoy, O. Y. Krakovetskiy, A. M. Zinchenko // Вісник Черкаського державного технологічного університету. Спецвипуск. – 2009. – С.11–13.
9. Краковецький О. Ю. Оцінка релевантності інформаційних блоків сайтів за допомогою SeoRank / В. М. Дубовой, О. Ю. Краковецький // Інформаційні технології та комп'ютерна інженерія. – 2010. – №1(17) . – С. 78–82.
10. Краковецький О. Ю. Очищення веб-сторінок від інформаційного шуму / В. М. Дубовой, О. Ю. Краковецький // Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2009): матеріали міжнародної конференції, м. Київ, 25-28 травня 2009 р. – С. 56–59.
11. Краковецький О. Ю. Отримання знань з експериментальних даних / В. В. Кабачій, О. Ю. Краковецький // Автоматика-2006: XIII міжнародна конференція з автоматичного управління, 25-28 вересня 2006 р. : тези доповіді. – Вінниця, 2006. – С. 411.
12. Свідоцтво про реєстрацію авторського права на твір. Україна. №22509. Комп'ютерна програма "Data Mining Studio" / О. Ю. Краковецький. Дата реєстрації: 31.10.2007.

АНОТАЦІЯ

Краковецький О. Ю. Інформаційна технологія оптимізації пошуку документів у веб-системах. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Вінницький національний технічний університет. – Вінниця. – 2011.

Дисертація присвячена розробленню інформаційної технології оптимізації пошуку докуметів у веб-системах з метою зменшення часу пошуку.

В роботі запропоновано метод оцінювання релевантності інформаційних блоків веб-сторінок з точки зору оптимізації для пошукових систем та метод очищення веб-сторінок від інформаційного шуму, який ґрунтується на розробленій математичній моделі оцінювання важливості інформаційних блоків веб-сторінок, що дало змогу підвищити ймовірність правильної ідентифікації основного контенту веб-сторінок. На основі цих методів та моделі було розроблено метод знаходження оптимальних шляхів перегляду документів у веб-системах, який дав змогу підвищити швидкість багатокрокового процесу отримання інформації. Розроблено методики та алгоритми для реалізації запропонованих математичних моделей і методів та на основі них розроблено програмне забезпечення.

Ключові слова: інформаційна технологія, автоматизована система управління, оптимізація для пошукових систем, евристичні алгоритми, оптимізація, інтелектуальний аналіз даних, асоціативні правила, кластеризація, регресійний аналіз.

АННОТАЦИЯ

Краковецкий А. Ю. Информационная технология оптимизации поиска документов в веб-системах. – Рукопись.

Диссертация на соискание учёной степени кандидата технических наук по специальности 05.13.06 – информационные технологии. — Винницкий национальный технический университет. — Винница. — 2011.

В работе проведено исследование методов информационного поиска, в частности, методов оценивания релевантности и полноты поиска, методов оценки дубликатов текстовых документов, а также методов определения основного контента веб-страниц. Также был проведен анализ проблем, с которыми сталкивается среднестатистический пользователь в процессе информационного поиска, который показал наличие целого ряда различных проблем. Исследованиями на тему улучшения методов и алгоритмов информационного поиска занимаются многие исследовательские центры и коммерческие поисковые компании, что свидетельствует об актуальности задачи оптимизации поиска документов в веб-системах.

В работе предложен метод SeoRank для оценивания релевантности информационных блоков веб-страниц с точки зрения оптимизации для поисковых систем, который основан на использовании принципов SEO, что позволило повысить достоверность идентификации типов информационных блоков с точки зрения важности для конечного пользователя. В отличие от классического применения правил SEO для комплексной оценки веб-сайтов, SeoRank не учитывает внешние факторы, а только те факторы, которые связаны с оценкой контента веб-страниц.

Получил дальнейшее развитие метод очищения веб-страниц от информационного шума, который основан на разработанной математической модели оценивания важности информационных блоков веб-страниц. Этот метод отличается тем, что учитывает ряд дополнительных семантических характеристик, которые позволили улучшить точность идентификации основного контента веб-страниц. В качестве тестовых данных для построения модели был использован корпус текстов компании Reuters, который насчитывает около 800 тыс. англоязычных текстов, а также дополнительные экспертные данные, собранные из современных информационных веб-сайтов.

В работе также предложен метод нахождения оптимальных маршрутов просмотра документов в веб-системах, который использует гипертекстовую структуру набора веб-документов для построения графа зависимости между ними, а также учитывает только основной контент веб-документов, что позволяет исключить из результатов поиска такие, которые содержат дублирующийся контент.

Метод нахождения оптимальных маршрутов просмотра документов в веб-системах лег в основу разработанной информационной технологии оптимизации поиска документов, которая позволяет уменьшить время поиска документов. Информационная технология не привязана к конкретному источнику данных, и, поэтому, может быть использована в любых информационных системах, где присутствует функция поиска и большое количество документом.

Разработаны методики практического применения предложенных методов и моделей, которые позволяют эффективно использовать полученные результаты для решения задач оптимизации поиска документов в веб-системах. В частности, разработанная информационная технология была успешно использована для повышения качества поиска в онлайн системе управления проектами, а также в функциональной подсистеме поиска корпоративного сайта, что позволило уменьшить время поиска в среднем на 25 %.

Разработанное алгоритмическое и программное обеспечение подтверждает адекватность и корректность теоретических выводов, а также практическую ценность результатов диссертационного исследования. Проведен сравнительный анализ разработанных методов с существующими методами, которые показали большую эффективность при решении задач информационного поиска.

Ключевые слова: информационная технология, автоматизированная система управления, оптимизация для поисковых систем, эвристические алгоритмы, оптимизация, интеллектуальный анализ данных, ассоциативные правила, кластеризация, регрессионный анализ.

THE SUMMARY

Krakovetskyi O. Y. Information technology for web search optimization. — Manuscript.

Thesis for achievement of a candidate’s degree on technical sciences on a specialty 05.13.06 – information technologies. — Vinnytsia National Technical University. — Vinnytsia. — 2011.

The thesis is devoted to development of information technology in search optimization Paper of Web-based systems to reduce the search time.

The method of evaluating the relevance of information blocks web pages in terms of optimization for search engines and method of cleaning the web of information noise, based on mathematical models assessing the importance of information blocks web pages, allowing us to

increase the likelihood of correct identification of the main content web pages. Based on these methods and models have been developed a method of finding optimal ways to view documents in Web-based systems that enabled to increase the speed of multistep process information. The appropriated algorithms and software were developed for proposed mathematical models and methods.

Keywords: information technology, computer control system, search engine optimization, heuristic algorithms, optimization, data mining, associative rules, clustering, regression analysis.

Підписано до друку 07.02.2011 р. Формат 29,7×42 1/4

Наклад 100 прим. Зам. № 2011-034
Віддруковано в комп'ютерному інформаційно-видавничому центрі
Вінницького національного технічного університету
Вінниця, вул. Хмельницьке шосе, 95. Тел.: 59-81-59