УДК 004.738.52

А.М. Пелещишин, П.І. Жежнич, Ю.О. Сєров
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ІНТЕЛЕКТУАЛЬНА СИСТЕМА АНАЛІЗУ СЛАБОСТРУКТУРОВАНИХ ВЕБ-РЕСУРСІВ

Розглянуто актуальну проблему створення інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів, призначеної для знаходження, видобування, структуризації інформації з слабоструктурованих джерел та подальшому аналізі видобутої інформації (проведення Web mining). Система складається з трьох підсистем: сканування Веб-ресурсів і потрібних даних, видобування структурованої інформації, аналітичної обробки інформації.

This paper considers information about Intelligent system Web-resource analyzing developing. Intelligent system Web-resource analyzing must solve problems of finding relevant documents, information extraction and informational retrieval. System consist of three subsystems: useful data finding and retrieving, structured information extraction, information analytic processing.

Постановка проблеми в загальному вигляді

Сьогодні інформація — одна з головних передумов успішного функціонування підприємств і організацій. Тому завдання збирання інформації для підприємств є надзвичайно важливим.

Щодня в Інтернеті з'являється багато (близько 7 мільйонів) нових сторінок. Оскільки більшість інформації, яка міститься в Інтернеті, є прихованою, то створення інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів є дуже актуальним завданням.

Значну частину прихованої інформації становлять тематичні бази даних: телефонні довідники, розклади руху транспорту, портали прогнозу погоди, а також сайти новин, форуми і т.д. Усі ці інформаційні ресурси є слабоструктурованими Веб-ресурсами.

Інтелектуальна система аналізу слабоструктурованих Веб-ресурсів покликана спростити користувачам доступ до таких джерел інформації шляхом структурування і аналізу даних, котрі містяться у таких тематичних базах даних.

Знаходження потрібної інформації є дуже складним завданням. Сучасні пошукові системи не забезпечують потрібного рівня індексації Веб-ресурсів, тому метою інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів є полегшення знаходження інформації та її подальшого аналізу.

Інтелектуальна система аналізу слабоструктурованих Веб-ресурсів допомагатиме користувачу знаходити та аналізувати дані, котрі містяться в мережі Інтернет, зокрема в тій частині, яка називається "прихований Веб" (hidden web).

Аналіз останніх досліджень

Web mining — це використання методів data mining для автоматизованого знаходження і видобування (extraction) інформації з Веб документів і сервісів. В цьому напрямку проводиться дуже багато досліджень в зв'язку з неймовірним ростом кількості інформації, яка з'являється в Веб і великою зацікавленістю в електронній комерції (e-commerce).

Web mining складається з таких підзадач:

1. Знаходження ресурсів – знаходження відповідних Веб-документів. Знаходження ресурсів — це процес знаходження даних і вибору її з текстових джерел, які містяться у Веб, таких як електронні новини, групи новин, текстове наповнення HTML з відкиданням теґів, а також ручна

вибірка Веб-ресурсів. До ресурсів також належать он-лайнові текстові ресурси, зроблені для дослідників, текстові бази даних тощо.

2. Вибірка(Selection) інформації і попередня обробка – автоматизована вибірка і попередня обробка інформації із знайдених Веб-ресурсів.

3. Узагальнення(Generalization) – автоматичне знаходження загальних закономірностей (pattern) як на окремих Веб-сайтах, так і на множині сайтів.

4. Аналіз – перевірка і/або інтерпретація знайдених закономірностей.

Web mining тісно пов'язаний з машинним навчанням і аналізом даних. Також Web mining часто асоціюється з пошуком інформації (Information Retrieval) та видобуванням інформації (Information Extraction), хоча насправді це не те саме.

Пошук інформації (Information Retrieval(IR)) — автоматичне знаходження усіх релевантних документів і водночас мінімізація нерелевантних документів серед знайдених, а також рангування знайдених документів за мірою релевантності. Видобування інформації (Information Extraction(IE)) полягає в обробці колекцій документів і наведенні інформації, яку вони містять, у формі, зручній для роботи і аналізування. На відміну від IR, метою якого є знаходження релевантних документів, метою IE є знаходження релевантних даних у документах.

Web mining поділяється на три категорії, відповідно до частин Веб, які можна досліджувати: дослідження вмісту Веб (Web content mining), дослідження структури Веб (Web structure mining), дослідження поведінки користувачів (Web usage mining).

Цілі статті

Призначення інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів полягає у знаходженні, видобуванні, структуризації інформації з слабоструктурованих джерел та подальшому аналізі видобутої інформації (проведення Web mining).

Відповідно до вимог, які ставляться перед системою, вона повинна складатися з таких компонент:

- підсистема сканування Веб-ресурсів і скачування потрібних даних (Crawler);
- підсистема видобування структурованої інформації із слабоструктурованих джерел інформації (Grabber);
- підсистема аналітичного опрацювання інформації.

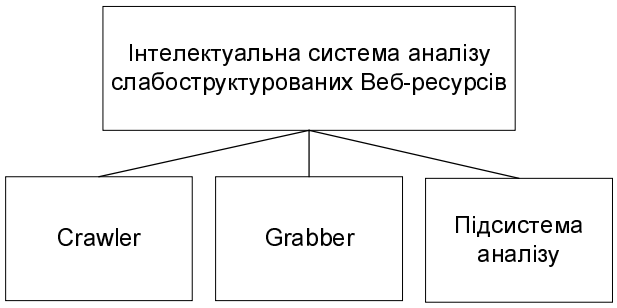

Рис. 1. Структура інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів

Підсистема сканування Веб-ресурсів і скачування потрібних даних призначена для сканування Веб-ресурсів, знаходженні потрібних ресурсів відповідно до заданих шаблонів і скачуванні знайдених ресурсів.

Підсистема видобування структурованої інформації із слабоструктурованих джерел інформації призначена для видобування потрібних даних зі знайдених ресурсів і їх занесення у базу даних відповідно до заданих правил.

Підсистема аналітичної обробки інформації призначена для проведення аналізу знайдених і структурованих ресурсів засобами data mining.

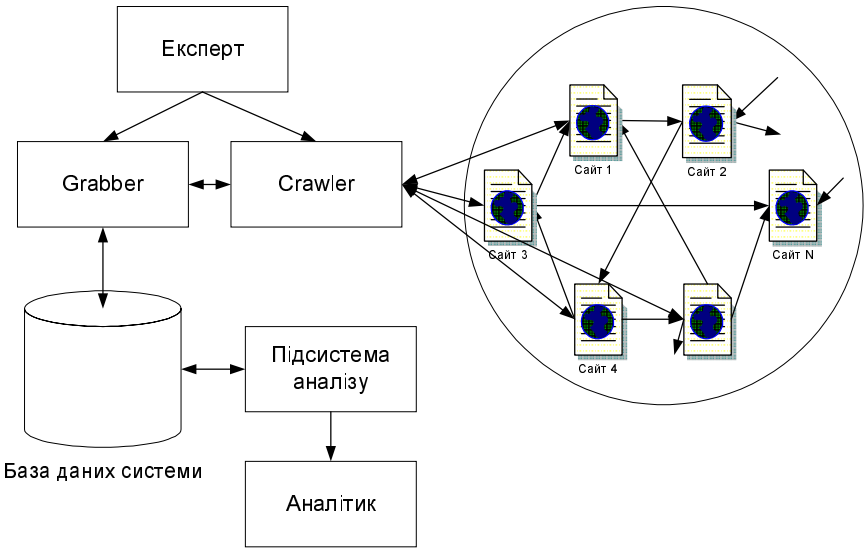

Рис. 2. Схема функціонування інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів

Основний матеріал

Застосування системи

Пошук інформації та її структурування потрібні всім: як великим підприємствам та організаціям, так і звичайним користувачам. Тому інтелектуальна система аналізу слабоструктурованих Веб-ресурсів може мати дуже широке застосування.

Інтелектуальна система аналізу слабоструктурованих Веб-ресурсів може застосовуватись для пошуку інформації про певних осіб, цін на товари у електронних магазинах; видобування даних з різноманітних тематичних баз даних (телефонні довідники, довідники руху транспорту, порталів новин і т.д).

Очікувані ефекти від впровадження системи

Результатом розробки інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів має стати повноцінний продукт, який буде виконувати завдання пошуку, скачування, структурування і аналізу інформації, котра міститься у непроіндексованій частині Вебу (hidden web).

Внаслідок впровадження системи інформація, котру неможливо сьогодні знайти за допомогою пошукових машин, стане доступною для користувачів.

Також інтелектуальна система аналізу слабоструктурованих Веб-ресурсів дозволить отримувати результат у структурованому вигляді, що полегшить її подальшу обробку та аналіз.

Це значно спростить і пришвидшить пошук інформації, дасть користувачу можливість на свій запит отримати комплексну інформацію, яка стосується об’єкта пошуку.

Описання функцій системи

Основним компонентами інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів є:

- Crawler — підсистема сканування Веб-ресурсів і скачування потрібних даних;
- Grabber — підсистема видобування структурованої інформації із слабоструктурованих джерел інформації;
- Підсистема аналізу видобутої інформації.

Опишемо функції кожної компоненти детальніше.

Crawler. Ця підсистема повинна виконувати наступні функції:

1. Сканування Веб-ресурсів і знаходження таких, що відповідають потрібним шаблонам
2. Скачування знайдених Веб-ресурсів.

Особливу складність становить реалізація функції сканування, тобто переходу з сайту на сайт, зі сторінки на сторінку. Необхідно розробити механізм реалізації комплексу правил, за якими б функція сканування здійснювала переходи за посиланням, які можуть нам дати корисний результат. Ігнорувати посилання на сторінки, на яких апріорі немає потрібної інформації і заходити лише на певні сторінки сайту.

Скачуваня. Функція скачування знайдених сторінок полягає у збереженні певних типів даних, які розміщені на Веб-сторінках і періодичному оновленні і доповненні скачаної інформації.

Grabber. Призначений для видобуваня структурованої інформації з слабо структурованих Веб-ресурсів. Видобування з *.html сторінок необхідних даних згідно з правилами видобування. Наприклад, видобування лише тієї інформації, що міститься в певних теґах.

Підсистема аналізу видобутої інформації. Призначена для здійснення аналізу видобутої інформації шляхом застосування методів data mining з метою отримання нових знань.

Описання вхідних і вихідних даних

Crawler. Вхідними даними для цієї підсистеми є:

- Шаблони структур сайтів і сторінок для сканування;
- Адреса сторінки, на яку потрібно зробити перехід;
- Правила скачування;

Вихідними даними є:

- Адреса сторінки, на які потрібно здійснити перехід;
- Адреси сторінок, які слід скачати.

Ця функція працюватиме рекурсивно, тому вихідний параметр “Адреса сторінки, на які потрібно здійснити перехід”, є вхідним для цієї функції на наступній ітерації.

Grabber. Вхідними даними підсистеми видобування структурованої інформації є:

- Шаблони даних, які слід видобути зі сторінок;
- Адреси сторінок, які слід скачати.

Вихідними даними є:

- Структуровані дані, видобуті зі сторінок.

Підсистема аналізу. Вхідними даними підсистеми аналізу є вихідні дані підсистеми Grabber:

- “Структуровані дані, видобуті зі сторінок.”

Вихідними даними є:

- Аналітичні звіти, які будуть результатом застосування data mining.

Формальна модель системи

Формальну модель системи подамо за допомогою діаграми потоків даних, яка зображена на рис. 3.

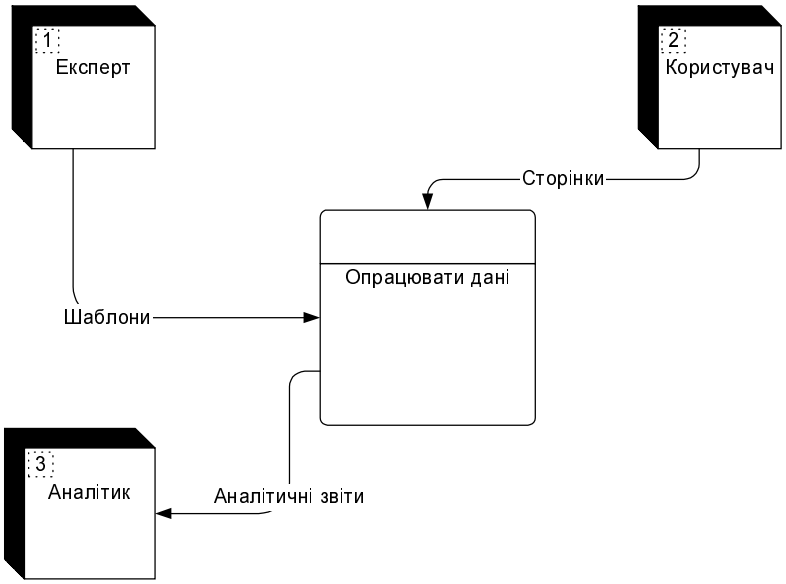

Рис. 3. Контекстна діаграма потоків даних “to be”

На рис. 4 зображено діаграму потоків даних “to be” першого рівня.

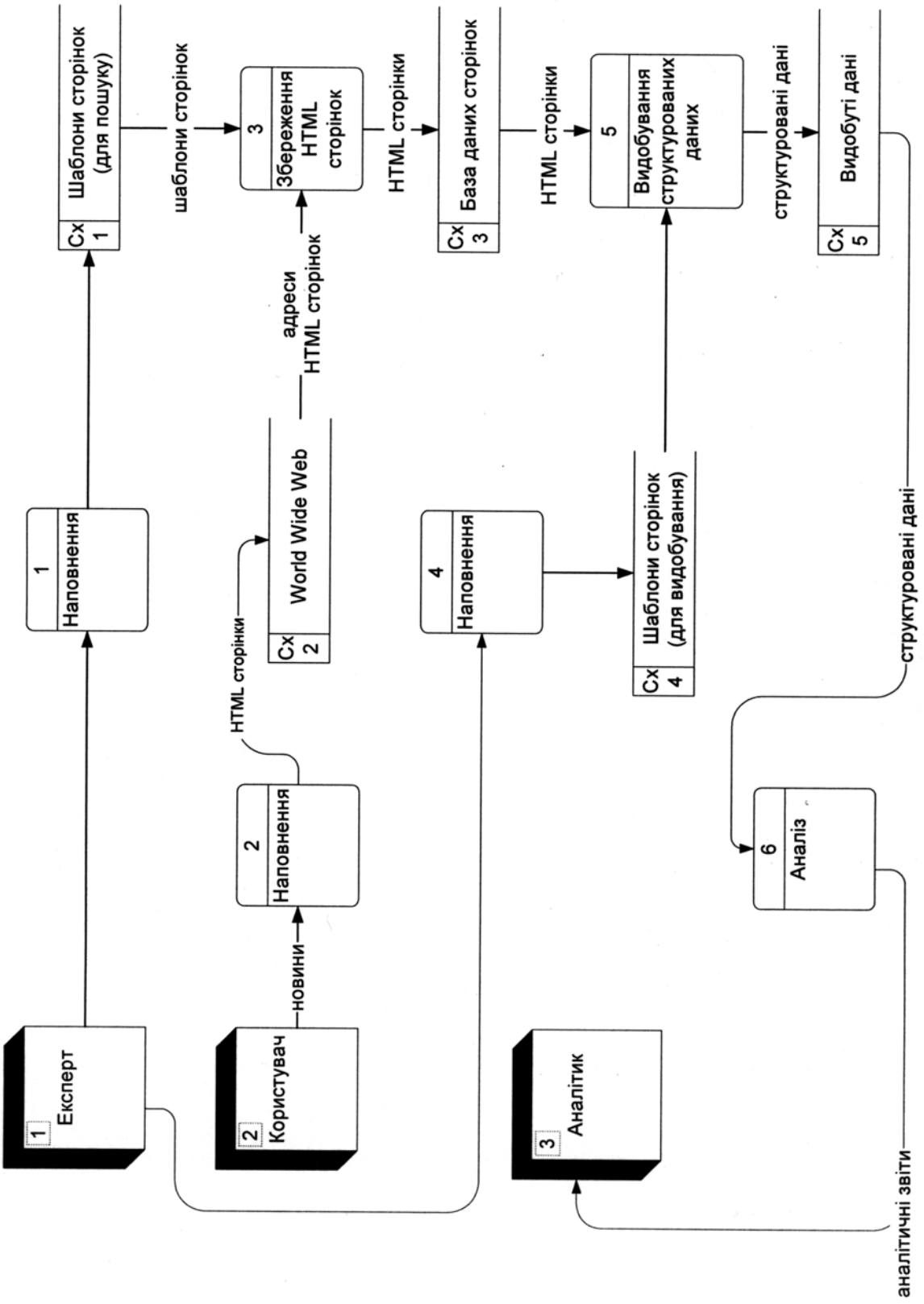

Рис. 4. Декомпозиція діаграми потоків даних “to be” першого рівня

Описання складу та структури бази даних

Опишемо структуру бази даних, у якій містяться дані, видобуті з Веб-сторінок.

ERD	Edit Date: 26.10.2003 14:47:56	
Description:		
Target DB: Access	Rev: 0	Creator: СєровЮрій
Filename: Paper.doc		Company: ICM

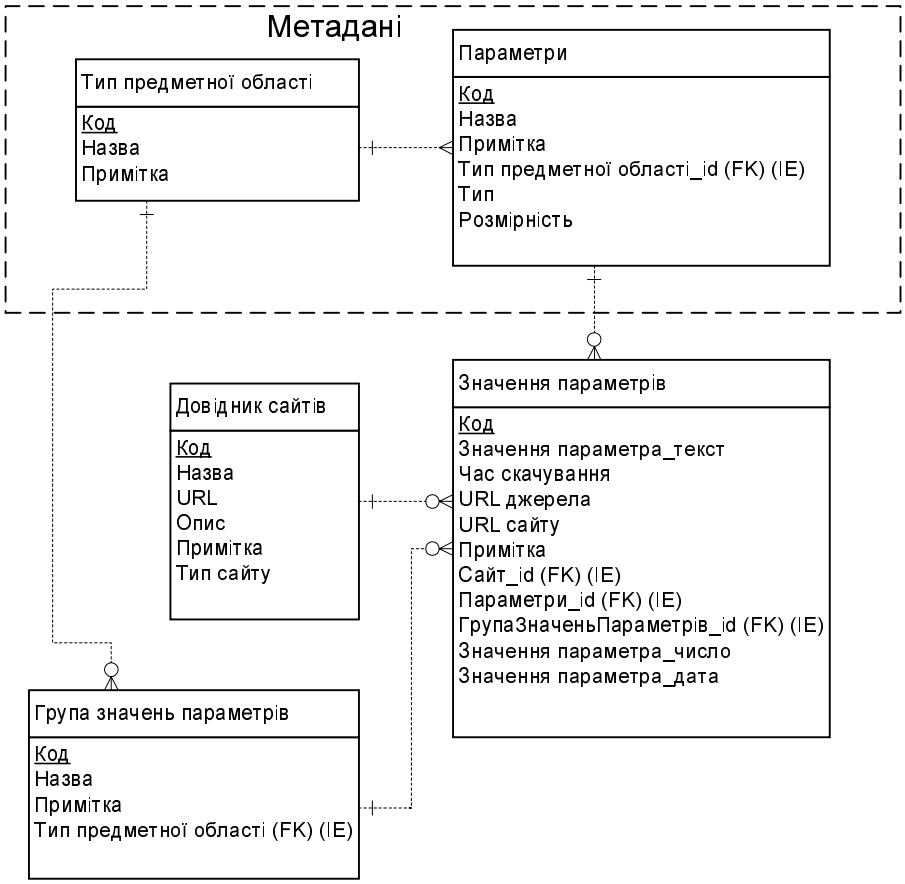

Рис. 5. Концептуальна схема бази даних

На рис. 5 зображена концептуальна схема база даних, представлена за допомогою діаграми “сутність–зв’язок”. Опишемо всі таблиці бази даних.

Таблиці “Тип предметної області” та “Параметри” призначені для збереження метаданих (інформація про дані, яка зберігатиметься у базі). Таблиця “Параметри” описує параметри предметної області. Для будь-якої предметної області можна задати потрібні параметри. Таблиця “Значення параметрів” призначена для зберігання значень параметрів, котрі описуються в таблиці “Параметри”. Таблиця “Довідник сайтів” зберігає дані про сайти, з яких видобуваються дані. Таблиця “Група значень параметрів” призначена для логічного об’єднання значень параметрів, котрі знаходяться у таблиці “Значення параметрів”. Бази даних для шаблонів сторінок для пошуку та збереження не створені, оскільки шаблони задаються у вигляді файлів.

Приклад застосування системи

Розглянемо приклад функціонування системи на прикладі сканування порталу прогнозів погоди, який знаходиться в Інтернеті за адресою www.accuweather.com.

Рис. 6. Портал прогнозу погоди Accuweather.com

На цій сторінці міститься прогноз погоди для міста Львів на період часу з 2 листопада по 7 листопада.

Нам не потрібні усі дані, які розташовані на сторінці, а лише ті дані, які стосуються погоди у Львові.

У результаті роботи програми видобування необхідних нам даних в базі даних буде містити таке:

Таблиця 1

Тип предметної галузі

Код	Назва	Примітка
000001	Погода	

Таблиця 2

Параметри

Код	Тип предметної області	Назва	Тип	Розмірність	Примітка
000001	Погода	Місто	Текст	100	
000001	Погода	Час	Дата/Час		
000001	Погода	Температура	Число	3	
000001	Погода	Вітер	Текст	50	
000001	Погода	Вологість	Число	3	
000001	Погода	Опади	Текст	50	

Таблиця 3

Значення параметрів

Код	Параметри	Значення	Час скачування	URL джерела	URL сайту	Примітка
00000001	Місто	Львів	01/11/03 15:17:23	http://www.accuweather.com/adcbin/public/intlocal_index.asp?reg=EU%3BEUROPE&cntry=EU%3BUR&wxcountry=EU%3BUR&wxcity2=L%27VIV+&Select+city=Submit&partner=accuweather	http://www.accuweather.com	
00000002	Час	02/11/03 12:00	01/11/03 15:17:23	//-//	//-//	
00000003	Температура	14	01/11/03 15:17:23	//-//	//-//	
00000004	Опади	Так	01/11/03 15:17:23	//-//	//-//	
00000005	Реальне відчуття	9				
00000006	Місто	Львів	01/11/03 15:17:23	//-//	//-//	
00000007	Час	03/11/03 00:00	01/11/03 15:17:23	//-//	//-//	
00000008	Температура	5	01/11/03 15:17:23	//-//	//-//	
00000009	Опади	Ні	01/11/03 15:17:23	//-//	//-//	
00000010	Реальне відчуття	-1	01/11/03 15:17:23	//-//	//-//	
…	…	…	…	…	…	…

Таблиця 4

Група значень параметрів

Код	Тип предметної області	Місто	Час	Температура	Опади	Реальне відчуття	Примітка
000001	Погода	Львів	02/11/03 12:00	14	Так	9	
000002	Погода	Львів	03/11/03 00:00	5	Ні	-1	
…	…	…	…	…	…	…	…

Висновки

У статті розглянуто актуальну задачу побудови інтелектуальної системи аналізу слабоструктурованих Веб-ресурсів.

Інтелектуальна система аналізу слабоструктурованих Веб-ресурсів призначена для вирішення проблем, які постають перед користувачами в ході пошуку необхідної інформації та її опрацювання:

- знаходження релевантної інформації;
- виведення знань з інформації, яка міститься у Веб;
- персоналізація інформації;
- знання про споживачів/користувачів.

Інтелектуальна система аналізу слабоструктурованих Веб-ресурсів складається з таких підсистем:

- сканування Веб-ресурсів і скачування потрібних даних (Crawler);
- видобування структурованої інформації із слабоструктурованих джерел інформації (Grabber);
- аналітичного опрацювання інформації.

Отже, система забезпечує вирішення основних задач, які постають перед користувачами в ході пошуку необхідної інформації, зокрема в його прихованій частині.

Система розроблялась на основі відкритих стандартів з використанням вільно поширюваного програмного забезпечення (Perl, MySQL).

1. Жежнич П.І., Кравець Р.Б., Пасічник В.В., Пелещишин А.М. Основні правила побудови семантично відкритих інформаційних систем // Вісник Національного університету "Львівська політехніка" "Інформаційні системи та мережі". – 1999. – №383. – С. 84–95. 2. Жежнич П.І., Кравець Р.Б., Пасічник В.В., Пелещишин А.М. Семантично відкриті інформаційні системи // Вісник Національного університету "Львівська політехніка". – 1999. – №383. – С. 73–84. 3. Сєров Ю. О. Технології пошуку та видобування даних у WWW (аналіз проблеми) // Вісник Національного університету "Львівська політехніка" "Інформаційні системи та мережі". – 2003. – №489. – С. 276–286 4. Web Mining Research: A Survey, www.cs.kuleuven.ac.be/~dtai/publications/files/33042.ps.gz Informational Retrieval on the Web, www.trl.ibm.co.jp /kobayashi00information.pdf Mining the Link Structure of the Web www.almaden.ibm.coman /chakrabarti99mining.pdf Journey to the Internet's Unknown Regions, http://www.newsfactor.com/perl/story/17418.html, 22.06.2003, Information Extraction from the Web, Wolfgang May, http://citeseer.ist.psu.edu/cache/papers/cs/18890/http:zSzzSzwww. informatik. unifreiburg.dezSz~dbiszSzPublicationszSz2KzSzTR136-InfoExtr.pdf/may00information.pdf, 20.09.2004, Effective Web Data Extraction with Standard XML Technologies, Jussi Myllymaki myllymaki01effective.pdf-http://citeseer.ist.psu.edu/cache/papers/cs/22164/http:zSzzSzwww.www10.orgzSzcdromzSzpaperszSzpdfzSzp102 .pdf/myllymaki01effective.pdf, 20.09.2004, A Brief Survey of Web Data Extraction Tools, Alberto H. F. Laender http://citeseer.ist.psu.edu/cache/papers/cs/26826/http:zSzzSzlsirpeople.epfl.chzSzabererzSzcitationsz Szsigrec02.pdf/laender02brief.pdf, 20.09.2004.

УДК 681.3

Пелещишин А.М, Корнилюк В.М.
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

КОМПЛЕКСНА ІНФОРМАЦІЙНА СИСТЕМА „СТУДМІСТЕЧКО”

Розроблено проект інтелектуальної системи підтримки прийняття рішень у сфері управління студентським містечком. Основною метою діяльності системи є забезпечення оперативного доступу до актуальної інформації про поточний стан студентського містечка, автоматизувати роботу із створення різноманітних звітів про діяльнысть студмістечка, надавати інформацію у вигляді, зручному для оперативного та стратегічного планування діяльності студмістечка. .

Project of the intellectual system of support of acceptance of decisions in the field of management of campus. Primary purpose of system activity there is providing an efficient access to the actual information about the current state of the campus, to automatize work on creation of various reports on the campus activity, to give information as comfortable for the efficient and strategic planning of campus activity.

Постановка проблеми в загальному вигляді

Студмістечка великих та середніх навчальних закладів (одним з яких є студмістечко Національного університету „Львівська політехніка”) часто охоплюють комплекс з більше ніж десятка гуртожитків, в яких можуть проживати десятки тисяч мешканців. Часто до складу