УДК 51.001.57+004.652.4+004.827

Н.Б. Шаховська

ОЦІНЮВАННЯ ЯКОСТІ КОНСОЛІДОВАНИХ ДАНИХ

Стаття присвячена опису особливостей оцінювання якості даних, отриманих з різних джерел, та розробленню алгоритму визначення релевантності відповіді користувачу.

Вступ

Сучасний рівень розвитку інформаційної технології (ІТ) все більше набуває глобалізаційного характеру. «Цінною» вважається інформація, яка отримана з різних джерел, подана під різними кутами зору, але водночас є цілісною та несуперечливою. Консолідованими даними називають системно інтегровані повні несуперечливі дані, придатні для підтримки прийняття рішень.

Внаслідок керування різнотипними даними з метою розв'язання аналітичних задач стратегічного рівня перед дослідниками виникає задача якості даних – відповідності вимогам користувачів. На рівні задач, для яких використовується точкове джерело, якість даних цього джерела є достатньою, і задовольняє (повністю чи частково) потреби осіб, що приймають рішення на їх основі. Проте використання даних з декількох джерел, наперед неузгоджених та з невідомими структурами, призводить до того, що якість даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне подання, необхідне для вирішення проблеми.

Під оцінюванням якості даних розумітимемо процес компонування даних, очищення та вдосконалення даних, а також об'єднання з усуненням дублювання та невизначеності. До якісних даних ставляться такі вимоги: повнота, точність, зв'язність, доступність, актуальність, своєчасність. Відсутність хоча б однієї з вищенаведених характеристик впливає на правильність рішення, прийнятого на основі консолідованих даних.

Реалізовані в сучасних серверах баз даних засоби аналізу та видобування даних (MS Analysis Server, Oracle Analytics тощо) не дають змоги враховувати наявність шуму, що, у свою чергу, породжує формування помилкових залежностей даних. Особливо ця проблема загострюється тоді, коли дані надходять з різних джерел, у тому числі і неструктурованих. На підтвердження цього ми спостерігаємо стрімке поширення NoSQL.

Однією з інформаційних технологій забезпечення опрацювання різнотипних інформаційних джерел даних є простір даних.

Дамо ряд визначень.

Інформаційний ресурс (ІР) Ir – сукупність даних в інформаційних об'єктах. Характеризується структурою даних Sd.

Інформаційний продукт (ІП) – документований інформаційний ресурс, який є результатом функціонування інформаційної технології $Ip = \langle Ir, Cg_{Ip}, Rl \rangle$, де Cg_{Ip} – каталог, Rl – методи доступу. Інформаційними продуктами є програмні засоби, текстові файли, веб-сторінки, електронні таблиці, xml-файли, бази даних, сховища даних, дані інших форматів.

Каталог ІП – метадані про ІП $Sd \cup Pl \cup Rl \to Cg_{Ip}$. Описує місцезнаходження ІП Pl, його структури даних (СДІР), методи доступу до ІР тощо.

Простір даних (ПД) – це блоковий вектор, що містить множину інформаційних продуктів предметної області, поділену на три блоки: структуровані дані *St* (бази, сховища даних), напівструктуровані дані *SemS* (XML, електронні таблиці) та неструктуровані дані *Ns* (текст). Над цим вектором та його окремими елементами визначено операції та предикати.

Сховище консолідованих даних cg' – віртуально побудоване сховище, що містить результат запиту користувача, отри-

маний з різних інформаційних продуктів шляхом узгодження структур даних.

1. Постановка проблеми в загальному вигляді

Проблемою якості даних займалися ще з 80-х років минулого століття. Так, Ванг, Кон сформували складники якісної інформаційної системи, а саме поняття «якість» трактували як відповідність встановленим вимогам користувача (рис. 1).

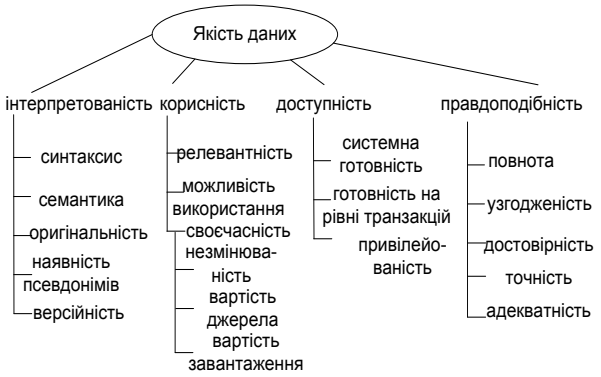

Рис. 1. Складові якості

Якість також означає повноту даних для ефективного прийняття рішень.

Загалом методи оцінювання якості в ІТ призначені для оцінювання якості вебсторінок (релевантності) та програмних засобів.

Модель RADCAB [1] має 6 основних параметрів для оцінювання якості веб- сторінок: релевантність (відповідність до теми пошуку) відповідність (чи доречна ця інформація користувачеві, наприклад, за віковою категорією), детальність (кількість інформації), часова характеристика (дата створення або оновлення), авторитетність (компетентність автора), нахил (причина створення документа). Передбачає експертне оцінювання, зведення оцінок експертів за допомогою методу аналізу ієрархій.

Аналогічним чином здійснюють оцінювання якості веб-сторінки Ciolek T.M. та Standler R.B. [2, 3]. Проте останній параметри якості зводить у функцію якості, експерти визначають вагу параметрів, і далі оцінюються ті веб-сторінки, у яких найбільше значення показника з найбільшою вагою.

На жаль, у просторі даних недостатньо лише експертного оцінювання якості даних через велику кількість джерел і необхідність реалізації пошуку метаданих.

Наступна група методів здійснює оцінювання інформативності. Так, індексний метод [4] використовується для оцінювання адекватності числових даних, здійснює згортку параметрів і використовується для визначення оптимізації запитів. Проте він не може бути використаний для визначення адекватності текстових даних. Метод Коваль Г.І. [5] використовується для оцінки якості ПЗ. За деякими змінами його можна було б застосовувати для визначення якості джерел даних, де явно вказано нижні та верхні межі надійності та відмовостійкості, підбір параметрів здійснюється експертно методом аналізу ієрархій. Проте для простору даних ця група методів також незастосовна через неможливість попереднього встановлення значення відмовостійкості певного джерела даних.

Наступна група методів для оцінювання якості використовує функцію корисності. Серед методів цієї групи доцільно виділити методи Згуровського М.З., Панкратової Н.Д. [6] та Соловйової К.О. [7]. Передбачають визначення корисності від додавання джерела, класифікування ситуацій прийняття рішень за рівнем невизначеності. Вводиться метрика якості рішення, прийнятого на основі заданих даних. Якщо після введення нового джерела якість даних знижується, приймається рішення про його видалення з простору даних. Проте необхідно вказувати межі інформативності джерел даних, що у випадку простору даних потребує доопрацювання, оскільки маємо справу з динамічною системою (з'являються нові джерела даних).

Метод корисності Афонічкіна А.І. [8] полягає у порівнянні якості прийнятих рішень на основі даних з невизначеністю та після усунення невизначеності. У випадку ПД це також необхідно робити у зв'язку з невизначеністю, що з'являється в консолідованих даних.

Отже, саме у напрямі формування функції якості консолідованих даних доцільно реалізувати оцінювання якості.

Метою статті є введення показників якості консолідованих даних, визначення функції якості та методу оцінювання якості даних.

2. Визначення показників якості даних

У просторах даних домінуючого значення набувають самі дані, їхнє зберігання і опрацювання. Для оцінювання якості даних застосуємо спільний методичний підхід до виділення адекватної номенклатури стандартизованих в ISO 9126 базових характеристик і субхарактеристик [9]. Базовими характеристиками стандарту є: функціональна придатність до використання; коректність або достовірність; ресурсна економічність; практичність; супроводжуваність; мобільність.

Функціональна придатність визначається, у першу чергу, повнотою накопичених об'єктів – відносною кількістю об'єктів або документів, наявних у джерелах даних, до загальної кількості об'єктів, що протралили у локальне сховище, яке містить консолідовані дані:

$$z_1 = |cg'| / |Ip_i.Ir| . \qquad (1)$$

Коректність або достовірність даних – це ступінь відповідності даних про об'єкти в базах даних реальним об'єктам у даний момент часу, що визначається змінами самих об'єктів чи їх характеристик. Визначена як відносна кількість описань об'єктів, які не містять дефектів і помилок, до загальної кількості об'єктів у просторі даних:

$$z_2 = |\sigma_P(Ip)| / |cg'| . \qquad (2)$$

Використовуваність ресурсів (або ресурсна економічність) у стандарті відображається зайнятістю ресурсів центрального процесора, оперативної, зовнішньої та віртуальної пам'яті, каналів введення-виведення, терміналів і каналів зв'язку. Цей показник у роботі не аналізується, оскільки існують розроблені методи (наприклад, метод критичних робіт) та засоби визначення завантаженості ресурсів.

Практичність – визначає застосовність консолідованих даних для певних користувачів. Оцінка практичності здійснюватиметься за допомогою функції корисності прийнятих рішень $v(r)$.

$$z_3 = v(r) / |cg'| .$$

Окрім того, цей показник враховує залежність прийнятого рішення від рівня довіри.

Супроводжуваність даних відображається зручністю і ефективністю виправлення, удосконалення або адаптації структури та змісту описань даних залежно від змін у зовнішньому середовищі застосування, а також у вимогах і функціональних специфікаціях замовника. Узагальнено якість супроводжуваності консолідованих даних можна оцінювати потребою ресурсів для її забезпечення і для реалізації. У просторах даних характеристика супроводжуваності пов'язана зі зміною даних про джерела даних у каталозі.

$$z_4 = |\sigma_{meta\_upd}(Ip_i.Cg)| / |cg'| . \qquad (3)$$

Мобільність характеризується тривалістю і трудомісткістю інсталяції інформаційних продуктів, адаптації та заміщення при перенесенні на інші апаратні та операційні платформи. Він також не використовується для оцінювання якості даних.

Отже, показниками якості даних є: функціональна придатність, коректність, практичність, супроводжуваність. Усі ці показники безрозмірні, $z_i \in [0...1]$, $i = \overline{1, 4}$.

3. Метод визначення якості даних

Визначимо корисність даних з ІП стосовно прийняття рішення на їх основі. Оцінку корисності даних здійснено наступним чином. Є множина керованих змінних $Z = (z_1, z_2, z_3, z_4)$. Визначено неперервну функцію якості Q. Допустима множина розв'язків замкнена, непуста й обмежена, оскільки за визначенням при-

наймні 2 показники якості (z_1, z_2) більші за нуль.

Цільова функція якості при обмеженнях має глобальний максимум:

$$Q(z_1,...,z_4)=\sum_{i=1}^{4}\left(\sum_k r_k z_i \prod_{j=1} P_{ij}\right)\to \max, \quad (4)$$

$$1\ge z_1\ge 0.75 \qquad z_1 c\le C_1 \qquad 1\ge z_4\ge 0.5$$
$$z_3\ge 0 \qquad z_4 c\le C_2 \qquad z_1 t_s\le T$$
$$0.25\ge z_1-z_2\ge 0 \qquad z_2 t_s\le T$$

де j вказує на інформаційний продукт, P_{ij} – рівень довіри до інформаційного продукту j для рішення k, r_k – оцінка рішення k, C_1 – загальна вартість завантаження об'єктів, C_2 – загальна вартість модифікації описів, T – загальний час завантаження, t_s – середній час завантаження одного об'єкта, c – середня вартість завантаження (модифікації) одного об'єкта.

Це задача нелінійної оптимізації з лінійними обмеженнями, яка вирішується певними методами (наприклад, градієнтний).

Поряд з фактичним оцінювання якості консолідованої інформації (4) необхідно провести оцінювання якості еталонного зразка, що відображає найкраще прийняте рішення. Потім виконується нормування фактичної оцінки, де k_i – ранг важливості, $k_i\in[0;1]$,

$$Q_i^e=\sum_i n_i z_i^e, \quad Q_{const}^{\prime e}=\sum_i k_i Q_i^e, \qquad (5)$$

$$Q_{const}=Q'_{const}/Q'^{e}_{const}, \qquad (6)$$

Для того, щоб підвести підсумок, наведемо основні етапи процедури оцінювання якості консолідованих даних.

Складання системи характеристик якості консолідованих даних. Ця система має вигляд ієрархічної структури. Для різних методик характерна різна кількість рівнів ієрархії, а також різна кількість критеріїв кожного рівня ієрархії. Система критеріїв якості може включати як внутрішні, так і зовнішні характеристики даних. Однак перевагу слід віддавати зовнішніми характеристиками. Крім того, критерії можуть носити як кількісний, так і якісний характер. Перевагу варто віддавати кількісним характеристикам. У нашому випадку мова йде про кількість показників Z.

Визначення значень відносних вагових коефіцієнтів $r_1,...,r_4$ характеристик якості із залученням думок експертів. Рекомендуємо здійснювати методом аналізу ієрархій. У випадку наявності різних думок – використовувати коефіцієнт конкордації Кендела.

Оцінювання значень показників якості $z_1,...,z_4$. Інформація про значення показників може бути отримана за результатами випробувань, експертного чи соціологічного опитування. Найкращим є перше джерело, але у випадку, якщо оцінка критеріїв цим методом неможлива або надмірно трудомістка, то залучається експертна інформація.

Нормування значень одиничних показників якості. У різних методиках використовуються різні функції приведення. Для пропонованої методики розрахунку значень показників нормування проводити недоцільно.

Обчислення факторів якості на підставі розрахунку зваженої згортки значень одиничних показників якості. У різних методиках використовуються різні оператори згортки і різне число кінцевих показників якості.

Далі розробимо алгоритм визначення відповідності рішення, прийнятого на основі консолідованих даних, еталонному.

Укрупнений алгоритм визначення відповідності рішення еталонному подано так.

1. Отримання параметрів вибірки еталонних та консолідованих даних.

2. Визначення критеріїв оптимальності.

3. Визначення найкращого значення за критерієм.

4. Визначення найгіршого значення за критерієм.

5. Пошук прямо пов'язаних даних.

6. Групування вибраних даних.

7. Обрання тих консолідованих даних, у яких агреговані кількісні характеристики рівні середньому значенню критеріїв 2) і 3).

8. Визначення джерела даних, з якого отримано інформацію, що задовольняє 7.

Ступінь співпадіння критерію з еталонним для заданих параметрів буде визначатись як

$$s = \sum_{i=1}^{n} a_i + a_t, \qquad (7)$$

де n – кількість нечасових параметрів співставлення; a_i – значення нечасового i-го параметра співставлення, яке набуває значення, a_t – значення часового параметра

$$a_i = \begin{cases} 0, \text{параметр не позначений} \\ 1, \text{параметр позначений} \end{cases};$$

$$a_t = \begin{cases} 0, \text{ігнорувати дати з джерел} \\ 1, \text{ігнорувати дати з еталону} \\ 2, \text{не враховувати інтервали} \\ 3, \text{перетин інтервалів} \\ 4, \text{повне співпадіння інтервалів} \end{cases}.$$

Слід зазначити, що кількість параметрів співставлення n буде різною для кожного типу рішення, що приймається. Обрання того чи іншого параметра фізично означатиме, що при співставленні знайдених даних та еталону за обраним атрибутом буде виконуватись операція агрегації для визначення релевантності. Чим більше параметрів буде включено до агрегування, тим точнішим буде отриманий результат співставлення. Обрання усіх параметрів означає максимальний ступінь довіри до отриманих результатів співставлення.

Можуть бути отримані такі результати співставлення ν:

– еталон не має аналога, $\nu = 0$;

– знайдені дані не відповідають жодному еталону, $\nu = 1$;

– часткове співпадіння (при агрегації даних еталону та знайдених даних отримано кількісні характеристики, які не рівні між собою) , $\nu = 2$;

– повне співпадіння, $\nu = 3$.

Блок схему алгоритму визначення ступеня співпадіння показано на рис. 2, 3.

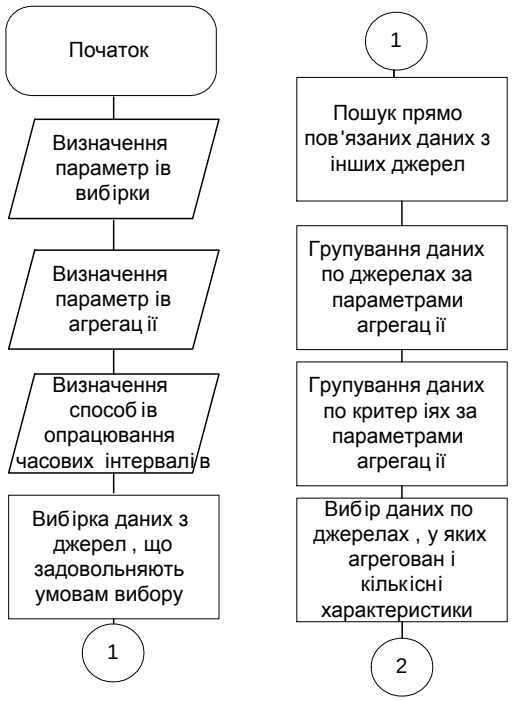

Рис. 2. Алгоритм визначення відповідності прийнятого рішення еталонному (початок)

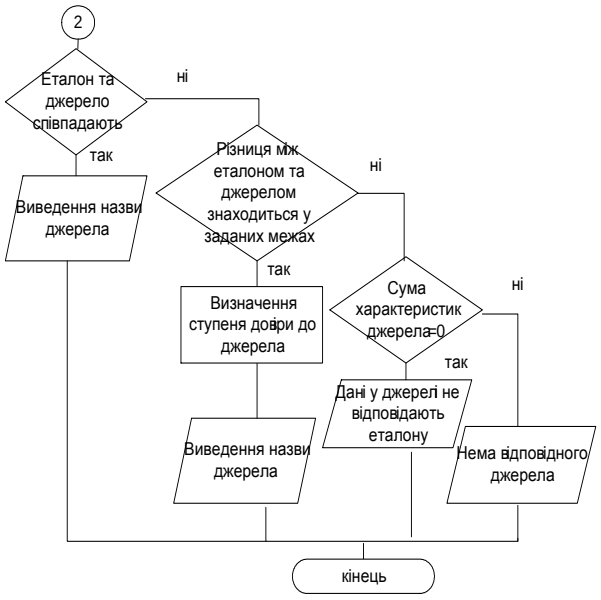

Рис. 3. Алгоритм визначення відповідності прийнятого рішення еталонному (закінчення)

4. Апробація методу

Тут досліджувалась якість консолідованих даних для текстових джерел даних. Результатом роботи є занесення у консолідоване сховище даних інформації з резюме пошукувачів роботи, причому резюме укладені у довільній формі. Ставилася задача аналізу резюме, де міститься наступна інформація:

1) дані про колишнє місце праці анкетованого;

2) дані про бажане місце праці анкетованого за такими характеристиками:

а) назва місця праці;

б) посада;

в) стаж роботи;

г) зарплата.

Є чотири множини, що не перетинаються: М – «Місце»; П – «Посада»; С – «Час»; З – «Зарплата». На початку роботи усі вони є порожніми:

М={ }; П={ }; С={ }; З={ }.

Подається набір значень

$$X = \{x_1, x_2, ..., x_n\};$$

n є числом цілим невід'ємним і відносно невеликим: $0 \le n \le 50$. Результатами виконання алгоритму є формування цих множин.

Метод, використаний у системі аналізу анкет, заснований на основі методу вибірки з множини документів. Вхідний текст поділяється на лексичні одиниці, для кожної з яких будуються класифікаційні правила визначених розмірів (у даній системі від 1 до 6 слів у правилі). Далі кожне з цих правил аналізується за правилами аналізу семантичних елементів, правилами порівняння слів, словником моделей тощо). Обираються ті класифікаційні правила (тобто набори слів), які згідно з результатами аналізу описують один з підпунктів (а, б, в, г).

Далі система аналізує текст загалом, щоб дослідити логічні зв'язки між обраними правилами. Наприклад, нехай запропонований текст: «Я рік працював в інституті викладачем. Потім моїм місцем праці була фірма, де я отримував 500 $ у місяць.». Після побудови та аналізу правил система видала б результат, наведений у табл. 1.

Таблиця 1. Результати роботи системи після аналізу ключових слів

Місце праці	Посада	Час	Зарплата
Інститут	Викладач	Рік	500 $ у місяць
Фірма			

Після аналізу тексту загалом система видасть результат, наведений у табл. 2.

Таблиця 2. Результати роботи системи після аналізу тексту

Місце праці	Посада	Час	Зарплата
Інститут	Викладач	Рік	–
Фірма	–	–	500 $ у місяць

Загалом система опрацювання анкет функціонує за схемою, показаною на рис. 4.

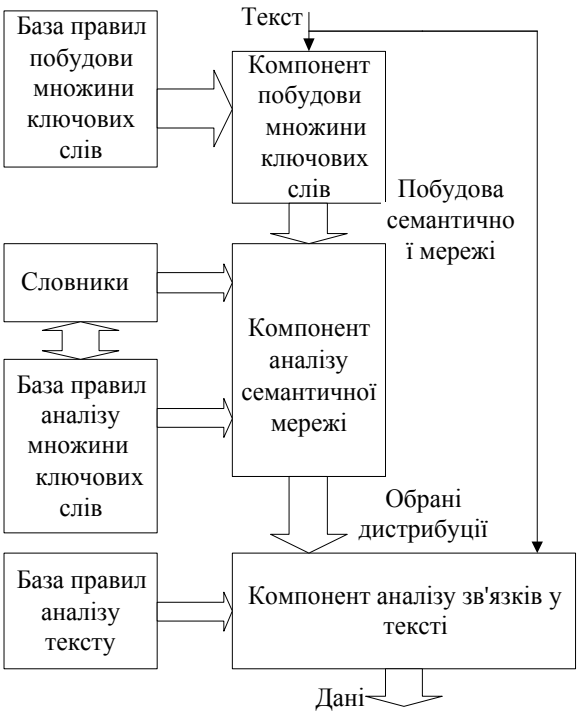

Рис. 4. Схема опрацювання анкет системою

Компонент побудови множини ключових слів, користуючись правилами побудови класифікаційних правил, здійснює розбиття тексту на лексичні одиниці, обирає слова, будує для них лексеми різних розмірів. Вхідними даними для нього є текст, результатом роботи – список лексем. Компонент аналізу класифікаційних правил аналізує кожну послідовність правил на наявність інформації про місце праці, обирає з них найбільш відповідні, потім процес повторюється для даних про посаду, час праці та зарплату. Вхідними даними для компоненту є список правил, результатом роботи – список правил для кожного з чотирьох питань. Компонент аналізу зв'язків у тексті, використовуючи правила аналізу тексту, впорядковує списки обраних правил за належністю до однієї події. Вхідними даними є список правил для кожного з чотирьох питань, вихідними – впорядковані дані по чотирьох питаннях.

Якість консолідованих даних перевірялася експертно. Результати наведені у табл. 3.

Таблиця 3. Результати консолідації текстових даних

Кількість анкет	% правильно визначених анкет	% частково визначених анкет	% неправильно визначених анкет
12	0,50	0,25	0,25
56	0,57	0,27	0,16
128	0,62	0,23	0,16
289	0,73	0,16	0,10
587	0,77	0,15	0,09

Чим більше анкет проаналізовано, тим точнішим є результат пошуку. Усунення невизначеності даних відбувається в сховищі консолідованих даних шляхом руху мережею записів. Аналіз результатів пошуку даних у джерелах наведено у табл. 4.

Таблиця 4. Результати оцінювання якості консолідованих даних

Відсоток правильних відповідей для	Пошук в ІП	Пошук у сховищі консолідованих даних, де	
		$z_1 > 0.9$	$z_2 > 0.9$
Посада	92	67	86
Місце роботи	93	91	92
Зарплата	87	74	81

Висновки

У статті проаналізовано методи визначення якості даних. *Наукова новизна*: розроблено метод визначення якості консолідованих даних на основі формалізації стандарту ISO 9126, що уможливило визначати придатність цих даних для подальшого прийняття рішень.

Практична цінність: розроблено засоби консолідації структурованих і неструктурованих даних та визначення їх якості, що дало змогу підвищити релевантність знайдених даних.

1. *Christensson K.* RADCAB– 2007. – http://www.radcab.com/about.html.
2. *Ciolek T.M.* Digitising Data on Eurasian Trade Routes: An Experimental Notation System – 2000. – www.ciolek.com/PAPERS /pnc-berkeley-02.html.
3. *Standler R.B.*, «Equations for Some Transient Overvoltage Test Waveforms System – 2004. –http://www.eeel.nist.gov/817/pubs/spd-anthology/files/Citations%20Part%204.doc
4. *Borisova E.* Index method of quality of the integrated complex objects – 1999. – http://www.mce.su/archive/doc15498/doc.pdf (in Russian).
5. *Koval H.* Models and methods of engineering quality software systems at the early stages of the life cycle: Kyiv: Kyiv national university Press, 2005 – 24 p.
6. *Zgurovskiy M., Pankratova N.* Basis of system analisys. – Kyiv: BHV, 2007. – 544 p. (in Ukranian).
7. *Sovovjova K.* Systemic and mathematical principles of natural classification and their use in intelligent systems. – Kharkov: Kharkov university of radioelectronics Press, 1999. – 34 p. (in Ukrainian).

8. *Aphonichkin A., Panphiloff A.* The quality of information provision in the management // Saratov: Saratov University Press, 1988. – 175 p. (in Russian).
9. *Shakhovska N.* Algebraic system of dataspace // Proc. of International Conference on Intellectual Systems for Decision Making and Problems of Computational Intelligence «ISDMCI'2011», 16–20 May 2011, Yevpatoria. – Vol. 1. – Kherson, 2011. – P. 14–18.

Одержано 10.06.2014

Про автора:

Шаховська Наталія Богданівна,
доктор технічних наук, доцент,
професор кафедри
інформаційних систем та мереж.

Місце роботи автора:

Національний університет
«Львівська політехніка»,
м. Львів,
вул. С. Бандери, 28.
Тел.: (032) 258 2404.
E-mail: natalya233@gmail.com