УДК 51.001.57

Р.Б. КРАВЕЦЬ, Ю.М. ОГРАДІНА

ФОРМАЛЬНІ ПІДХОДИ ДО МОДЕЛЮВАННЯ СИСТЕМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

На сучасному етапі розвитку інформаційних технологій великої актуальності набули розроблення методів побудови та організації інтелектуальних інформаційних систем (ІІС) і застосування цих систем для підтримки прийняття управлінських рішень. Такі системи повинні виступати єдиним джерелом інформації про предметну область (ПО) і поєднувати у собі цілий спектр засобів для отримання, збереження та опрацювання даних і знань про цю ПО. Однією з функцій ІІС, призначених для опрацювання великих масивів інформації, є набуття знань на основі даних, що зберігаються у базі даних системи. Для реалізації цієї функції за останній час створено технологію аналітичного опрацювання інформації (OLAP) [1], яка включає в себе засоби опрацювання даних на різних рівнях їх узагальнення, та інтелектуального аналізу даних (knowledge discovery in databases) [2], призначенням якого є віднаходження знань у базі даних. На сьогодні в дослідженнях інтелектуального аналізу даних (ІАД) розроблено ряд методів та алгоритмів, які дозволяють виявляти закономірності у даних і будувати моделі, що показують такі закономірності. Зауважимо, що ці методи розроблялися для побудови окремих видів моделей, що описують залежності у даних, найпоширенішими з яких є [3], [4]: класифікація, регресія, прогнозування, кластеризація, асоціація та послідовність. Побудова кожної з цих моделей формулюється як окрема задача, використовуючи при цьому власні формалізми, поняття та позначення [5], що унеможливлює формування єдиного підходу до вирішення задачі ІАД та створення систем ІАД як цілісного набору методів.

Виходячи із сказаного вище, виникає першочергова необхідність:

– створення формальної моделі системи інтелектуального аналізу даних (СІАД), визначення структури СІАД та її зв'язку з іншими видами інформаційних систем;

– формальної постановки задачі ІАД та вироблення єдиного підходу до її розв'язання.

У статті пропонується використання підходів загальної теорії систем [6] для побудови моделі СІАД та теорії реляційних баз даних [7] для математичної постановки задачі ІАД.

1. Модель та структура системи інтелектуального аналізу даних

В основі будь-якої ІІС лежить база даних (БД) та база знань (БЗ). На цьому етапі даними будемо називати конкретні значення, що описують властивості, характеристики та співвідношення сутностей ПО. Знанням вважатимемо деяку модель, що описує закономірності у наборах даних та взаємозв'язки, які виявляються між наборами даних. Сукупність даних та знань про ПО будемо називати інформацією про цю ПО.

Отже, інтелектуальна інформаційна система формально задається як трійка $\langle DB, KB, OP \rangle$, де DB – схема БД, KB – схема БЗ, OP – набір операцій над множинами DB та KB.

Схеми БД та БЗ конкретної ІІС визначаються на етапі проектування цієї системи. Набір операцій визначає, які дії можна виконувати над даними та знаннями у системі, а, отже якого типу ІІС можна побудувати, використовуючи наявні у наборі операції.

Розглянемо детальніше типи операцій над даними та знаннями в ІІС. У загальному виділимо операції у наступні групи.

– Операції над даними

До цієї групи належать операції, які опрацьовують дані, що зберігаються в БД, або змінюють стан БД, а саме: внесення первинних даних у БД, вибірка первинних даних із БД, вибірка опрацьованих даних із БД та ін. Зауважимо при цьому, що операції, які опрацьовують дані, не використовують знань, що зберігаються в БЗ інформаційної системи.

Позначатимемо набір операцій цієї групи через Op_1, а самі операції як

$$op_1 : Db \to Db$$

– Операції над знаннями

До цієї групи належать операції, що опрацьовують знання, які зберігаються у БЗ, або змінюють стан БЗ, а саме: внесення знань про ПО в БЗ, вибірка знань із БЗ, формування нових знань на основі

тих, що зберігаються у БЗ. Зауважимо, що до формування нових знань не залучаються дані, які зберігаються у БД ІС.

Позначатимемо набір операцій цієї групи через Op_2, а самі операції мають вигляд

$$op_2 : Kb \to Kb.$$

– Операції формування знань із даних

До цієї групи належать операції, які на основі даних із БД формують нові знання, що моделюють закономірності між цими даними. При цьому для видобування знань можуть використовуватися апріорні знання, що вже містяться у БЗ ІС. Тому операції цієї групи мають вигляд

$$op_3 : Db \times Kb \to Kb.$$

Набір операцій цієї групи позначатимемо як Op_3.

– Операції виведення нових даних

До групи операцій виведення нових даних входять операції, які на основі знань, що зберігаються в БЗ ІС, та даних із БД генерують нові дані, яких ще немає в БД. У загальному операції цієї групи мають вигляд

$$op_4 : Db \times Kb \to Db$$

Набір операцій виведення нових даних позначатимемо як Op_4.

Набір операцій, який складається з операцій над даними, операцій над знаннями, операцій формування нових знань та операцій виведення нових даних, тобто $Op = Op_1 \cup Op_2 \cup Op_3 \cup Op_4$, будемо називати *повним*.

У загальному випадку функції інформаційної системи можуть будуватися на основі деякої підмножини повного набору операцій (таблиця). Якщо така підмножина є типовою для багатьох прикладних інформаційних систем, то виділятимемо ці системи в окремий клас.

Види інформаційних систем з точки зору наборів операцій, на основі яких вони будуються

Система	Призначення	Операції
Система оперативного опрацювання інформації (OLTP-система)	Накопичення та вибірка первинних даних	Група Op_1 (частково)
Інформаційно-пошукова система	Вибірка первинних даних	Група Op_1 (частково)
Система аналітичного опрацювання інформації (OLAP-система)	Вибірка та опрацювання даних на різних рівнях агрегації	Група Op_1 (частково)
Експертна система	Накопичення знань експертів; виведення нових даних на основі знань та існуючих даних	Групи Op_2, Op_4,
Система підтримки прийняття рішень	Виведення нових даних на основі знань та існуючих даних	Групи Op_1, Op_2, Op_4
Система ІАД	Формування знань на основі даних	Групи Op_1 (частково), Op_3, Op_4 (частково)

На рис. 1 зображена компонентна структура ІС з повним набором операцій, показано зв'язок кожної компоненти з БД та БЗ, а також виділена СІАД.

Системою інтелектуального аналізу даних називатимемо ІС IS_{KDD} з набором операцій Op_{KDD}.

Набір Op_{KDD} включає такі операції:

– Операції попереднього опрацювання даних, а саме: поповнення, дискретизації, побудови концептуальної ієрархії, виділення основних факторів; ці операції належать до групи операцій над даними;

– Операції аналітичного опрацювання інформації, запропоновані у роботі [8]; що належать до групи операцій над даними;

– Операції видобування знань (data mining), які включають пошук асоціативних правил, класифікаційних правил та інші методи; ці операції належать до групи операцій формування знань із даних;

– Операція перевірки знань на відповідність даним, яка належить до групи операцій над знаннями

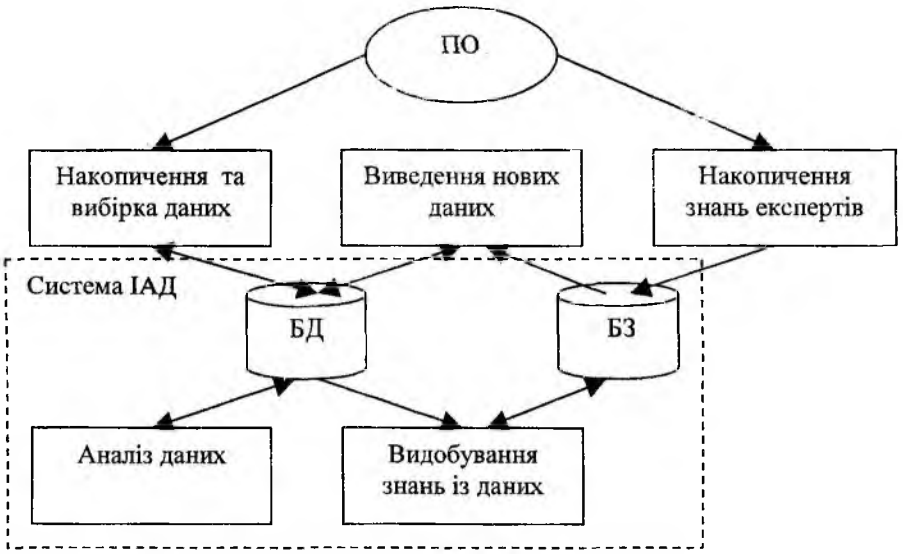

Рис. 1

На рис. 2 показано зв'язок СІАД з іншими видами інформаційних систем.

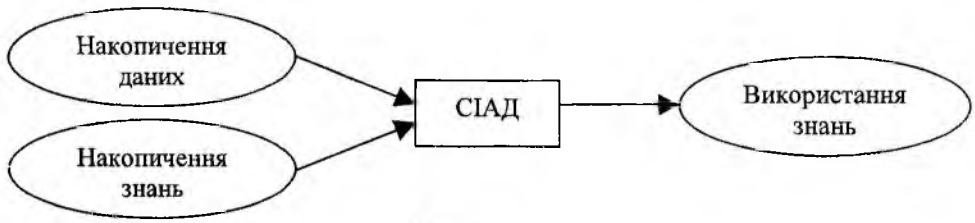

Рис. 2.

Призначенням СІАД є виявлення закономірностей та залежностей між даними про об'єкти, факти та події деякої ПО та збереження їх у формі знань у БЗ системи. Отже, основною функцією СІАД є формування знань про ПО на основі даних, що описують цю ПО. Для подальшого розгляду моделі СІАД виконаємо математичну постановку задачі ІАД.

2. Формальна постановка задачі інтелектуального аналізу даних

Основними елементами будь-якої ПО є її об'єкти, які, власне, і формують ПО як таку. У будь-який момент часу стан ПО повністю визначається сукупністю усіх її об'єктів. У свою чергу кожен об'єкт ПО має деякий набір властивостей, за допомогою яких описується об'єкт і які дозволяють осягнути суть об'єкта на потрібному спостерігачеві рівні абстракції.

Властивості об'єктів ПО називаються атрибутами. Кожен атрибут набуває значень із деякої множини (області визначення атрибута) – домену атрибута. Множину атрибутів ПО позначатимемо

$$U = \{A, dom(A)\}.$$

Кожен об'єкт ПО описується набором атрибутів з множини U. ПО ділиться на класи однотипних об'єктів, які описуються однаковими наборами атрибутів. Структуру класу об'єктів будемо позначати як

$$OB = \{Id_1,...,Id_k, A_1,..., A_m\},$$

де $ID = \{Id_i, dom(Id_i)\}_{i=1}^{k} \subset U$ – набір атрибутів, за значеннями яких однозначно визначається об'єкт ПО, і який будемо називати ідентифікатором об'єкта; $\{A_i, dom(A_i)\}_{i=1}^{m} \subset U$ $\{A_i, dom(A_i)\}_{i=1}^{m}$ – атрибут об'єкта ПО.

Об'єкт класу OB будемо позначати як ob, множину об'єктів цього класу як Ob, тобто $Ob = \{ob$.

Між об'єктами ПО можуть утворюватись зв'язки, які виникають, коли один об'єкт використовує властивості іншого. У такому випадку будемо говорити, що між об'єктами існує співвідношення.

Співвідношенням між об'єктами (ПО) ob_1, ob_2 зі структурами OB_1, OB_2 називатимемо впорядковану пару $(ob_1, ob_2)_{ob}$. Структури об'єктів задають структуру співвідношення між об'єктами як пару $(OB_1, OB_2)_{ob}$.

Серед усіх співвідношень окремо виділимо такі два типи зв'язків між об'єктами: співвідношення "частина–ціле" (мереономічна класифікація) та співвідношення "рід–вид" (таксономічна класифікація).

Співвідношення між об'єктами типу "частина–ціле", у якому об'єкт ob_1, зі структурою OB_1 є частиною об'єкта ob_2 зі структурою OB_2 будемо називати p-співвідношенням і позначати $(ob_1, ob_2)_{ob}^{p}$, а відповідну структуру співвідношення – $(OB_1, OB_2)_{ob}^{p}$.

Таксономічне співвідношення між об'єктами, у якому об'єкт ob_1 зі структурою OB_1 є видом об'єкта ob_2 зі структурою OB_2 будемо називати s-співвідношенням і позначати $(ob_1, ob_2)_{ob}^{5}$, а відповідну структуру співвідношення – $(OB_1, OB_2)_{ob}^{5}$.

Співвідношення між об'єктами ПО можуть утворювати ланцюжки, у яких об'єкт з одною структурою входить в одне співвідношення першим членом пари, а в інше – другим. У такому випадку отримаємо ієрархію об'єктів.

Ієрархією $h_{об}$ об'єктів (ПО) $ob_1,..., ob_n$ зі структурами $OB_1,...,OB_n$ називатимемо множину співвідношень між цими об'єктами виду $\{(ob_i, ob_{i+1})_{ob}\}_{i=1}^{n-1}$. Структура ієрархії задається як множина структур співвідношень $H_{ob} = \{(OB_i, OB_{i+1})_{ob}\}_{i=1}^{n-1}$.

Як і для співвідношень, окремо виділимо два типи ієрархій. Ієрархію, утворену p-співвідношеннями, будемо називати p-ієрархією (позначатимемо h_{ob}^{p} зі структурою H_{ob}^{p}), а ієрархію, утворену s-співвідношеннями, називатимемо s-ієрархією (позначатимемо h_{ob}^{5} зі структурою H_{ob}^{5}).

Елементами ПО, які визначають її розвиток в часі, є події. Подія ПО відображає зафіксовану в часі взаємодію між об'єктами ПО.

Однотипні події, тобто ті, що мають однакові атрибути і в яких беруть участь однотипні об'єкти, об'єднуються у класи. Структуру класу подій позначатимемо

$$EV = \{Id, Id_1^1,..., Id_{k_1}^1,..., Id_1^n,..., Id_{k_n}^n, A_1,..., A_m ,$$

де $Id \in U$ – ідентифікатор події, який приймає значення з множини моментів часу T; $ID' = \{Id_j^i, dom(Id_j^i)\}_{j=1}^{k_i} \subset U$ – ідентифікатор -го об'єкта; $\{A_i, dom(A_i)\}_{i=1}^{m} \subset U$ – атрибути події.

Події класу EV будемо позначати як ev, множину подій цього класу як Ev, тобто $Ev = \{ev\}$.

Між подіями ПО можуть утворюватись зв'язки. У такому випадку будемо говорити, що між подіями існує співвідношення.

Співвідношенням між подіями (ПО) ev_1, ev_2 зі структурами EV_1 та EV_2 називатимемо впорядковану пару $(ev_1, ev_2)_{ev}$. Структури подій задають структуру співвідношення між подіями як пару $(EV_1, EV_2)_{ev}$.

Окремо виділимо типи співвідношень між подіями, які виникають у таких випадках:

– поява однієї події чи подій певного класу спричиняють подію чи події того ж або іншого класу (причинно-наслідковий зв'язок між подіями);

– події одного класу використовують властивості подій іншого класу (таксономічна класифікація подій);

– події одного класу є складовими частинами подій іншого класу (мереологічна класифікація подій).

Причинно-наслідкове співвідношення між подіями, у якому подія ev_1 зі структурою EV_1 викликає подію ev_2 зі структурою EV_2, будемо називати q-співвідношенням і позначати $(ev_1, ev_2)^q_{ev}$, а відповідну структуру співвідношення – $(EV_1, EV_2)^q_{ev}$.

Таксономічне співвідношення між подіями, у якому подія ev_1 зі структурою EV_1 є видом події ev_1 зі структурою EV_2 будемо називати s-співвідношенням і позначати $(ev_1, ev_2)^5_{ev}$, а відповідну структуру співвідношення – $(EV_1, EV_2)^5_{ev}$.

Мереономічне співвідношення між подіями, у якому подія ev_1 зі структурою EV_1 є складовою події ev_2 зі структурою EV_2 будемо називати p-співвідношенням і позначати $(ev_1, e_2)_e$, а відповідну структуру співвідношення – $(EV_1, EV_2)^p_{ev}$.

Таксо- та мереономічні співвідношення між подіями ПО можуть утворювати ланцюжки, у яких подія з одною структурою входить в одне співвідношення першим членом пари, а в інше – другим. У такому випадку отримаємо ієрархію подій.

Ієрархією h_{ev} подій (ПО) ev_1, ..., ev_n зі структурами EV_1, ..., EV_n називатимемо множину s-співвідношень між цими подіями виду $\{(ev_1, ev_{i+1})_{ev}\}_{i=1}^{n=1}$. Структура ієрархії задається як множина структур співвідношень $H = \{(EV_i, EV_{i+1})_{ev}\}_{i=1}^{n-1}$.

Як і для співвідношень, будемо окремо виділяти два типи ієрархій. Ієрархію, утворену p-співвідношеннями, будемо називати p-ієрархією (позначатимемо h^p_{ev} зі структурою H^p_{ev}), а ієрархію, утворену s-співвідношеннями – s-ієрархією (позначатимемо h^5_{ev} зі структурою H^5_{ev}).

Причинно-наслідкові співвідношення між подіями дають можливість описати послідовності подій ПО.

Послідовністю sq подій (ПО) ev_1, ..., ev_n зі структурами EV_1, ..., EV_n називатимемо множину q-співвідношень між цими подіями виду $\{(ev_1, ev_{i+1})_{ev}\}_{i=1}^{n=1}$. Структура послідовності задається як множина структур співвідношень $SQ = \{(EV_i, EV_{i+1})^q_{ev}\}_{i=1}^{n-1}$.

Структура ПО повністю визначається множиною атрибутів, множиною моментів часу, структурами усіх об'єктів, співвідношень між об'єктами, подій та співвідношеннями між подіями. Таким чином, сукупність структур об'єктів, співвідношень і подій ПО утворюють структуру ПО:

$$Strukture(P) = \langle U, T, \{OB\}, \{H_{ob}\}, \{EV\}, \{H_{ev}\}, \{SQ\} \rangle.$$

Кожен одиничний об'єкт ПО описується даними, які будемо називати даними про об'єкт.

Означення 1. Даними про об'єкт (ПО) зі структурою $OB = \{Id_1, ..., Id_k, A_1, ..., A_m$ називається кортеж $ob = \{id_1, ..., id_k, a_1, ..., a_m\}$, де $id_i \in dom(Id_i)$ – значення атрибута-ідентифікатора Id_i об'єкта, $a_i \in dom(A_i)$ – значення атрибута A_i об'єкта.

Таким чином дані про об'єкти одного класу складають множину: $Ob = \{\langle id_1,...,id_k,a_1,...,a_m\rangle\}$.

Кожна одинична подія ПО описується даними, які будемо називати даними про подію.

Означення 2. Нехай T – деяка множина моментів часу; ob_1, ..., ob_n – дані про об'єкти ПО. Даними про подію (ПО) зі структурою $EV = \{Id, Id_1^1,...,Id_{k_1}^1,...,Id_1^n,...,Id_{k_n}^n, A_1,...,A_m$ будемо називати кортеж $ev = \{id, id_1^1,...,id_{k_1}^1,...,id_1^n,...,id_{k_n}^n, a_1,...,a_m$, де $id \in T$ – значення атрибута-ідентифікатора події; $id_j^i \in dom(Id_j^i)$ – значення атрибута-ідентифікатора Id_j^i об'єкта ob_i, $a_i \in dom(A_i)$ – значення атрибута події.

Таким чином дані про події одного класу складають множину:

$$Ev = \{id, id_1^1,...,id_{k_1}^1,...,id_1^n,...,id_{k_n}^n, a_1,...,a_m .$$

Сукупність усіх об'єктів, співвідношень між об'єктами, подій та співвідношень між подіями предметної області P повністю задають ПО протягом усього часу її існування.

Означення 3. Даними про ПО P будемо називати кортеж $P = \langle\{dom(A)\},\{Ob\},\{Ev\}\rangle$, де $\{dom(A)$ – множина доменів усіх атрибутів з множини U, $\{Ob\}$ – множина об'єктів ПО, $\{Ev\}$ – множина подій ПО.

На відміну від даних, які несуть у собі інформацію про конкретні об'єкти та події ПО, знання описують закономірності, що виникають всередині ПО, а саме: закономірності між властивостями об'єктів одного класу, співвідношення між об'єктами різних класів, закономірності між властивостями подій одного класу, співвідношення між подіями різних класів, закономірності у характері протікання процесів (ланцюжки подій як одного, так і різних класів) і т. п.

При формальному поданні знання будемо враховувати той факт, що дані, на основі яких отримуються ці знання, є структурованими і зберігаються у реляційній БД.

Розрізнятимемо знання для різних елементів ПО.

– Знання про властивість об'єктів одного класу

Нехай Ob – підмножина об'єктів одного класу зі структурою OB ; нехай $A \in OB$ – атрибут об'єктів цього класу, визначений на домені $dom(A)$. Знанням про атрибут A на множині об'єктів Ob будемо називати відношення $A \in \{a\}$, де $\{a\} \subset dom(A)$, якщо це відношення виконується для всіх об'єктів з множини Ob.

– Знання про сукупність властивостей об'єктів одного класу

Нехай Ob – підмножина об'єктів одного класу зі структурою OB; $\{A_i\} \subseteq OB$ – підмножина атрибутів об'єктів класу OB. Знанням про атрибути $\{A_i\}$ на множині об'єктів Ob будемо називати кортеж виду $\langle A_i \in \{a\}_1,..., A_k \in \{a\}_k\rangle$, де $\{a\}_i \subset dom(A_i)$, якщо усі відношення цього кортежу одночасно виконуються для усіх об'єктів з множини Ob.

– Знання про клас об'єктів

Нехай Ob – підмножина об'єктів одного класу зі структурою OB. Знанням про клас об'єктів OB на множині Ob будемо називати кортеж

$$\langle A_{i_1} \in \{a\}_{i_1},..., A_{i_k} \in \{a\}_{i_k}\rangle \to \langle A_h \in \{a\}_{i_1},..., A_{i_l} \in \{a\}_{j_l}\rangle,$$

де $\{a\}_i \subset dom(A_i)$, якщо відношення, задане цим кортежем, виконується для всіх об'єктів з множини Ob.

Аналогічно вводяться знання про властивості та класи подій.

Тепер можемо сформулювати постановку загальної задачі ІАД.

Нехай $P = \langle \{dom(A)\}, \{Ob\}, \{Ev\} \rangle$ – дані про ПО зі структурою

$$Strukture(P) = \langle U, T, \{OB\}, \{H_{ob}\}, \{EV\}, \{H_{ev}\}, \{SQ\} \rangle$$

Потрібно знайти знання про класи об'єктів та подій, а також співвідношення між об'єктами та співвідношення між подіями ПО на основі підмножин даних про ПО.

Задачі побудови моделей, які сьогодні розглядаються в ІАД, можуть бути сформульовані як частинні випадки загальної задачі ІАД. Зауважимо, що ці задачі охоплюють лише частину випадків загальної задачі ІАД.

3. Висновок

Запропоновані у статті формальні підходи до моделювання система інтелектуального аналізу даних є основою для застосування існуючих та побудови нових методів та алгоритмів видобування знань із баз даних і побудови систем інтелектуального аналізу даних

Список літератури: 1. *Codd E.F., Codd S.B., Salley S.T.* Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate // E.F. Codd & Associates. 1993. 2. *Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.* Advances in knowledge discovery and data mining. Mehlo Park: AAAI/MIT Press, 1996. 3. *Буров К.О.* Обнаружение знаний в хранилищах данных // Открытые системы. 1999. №5-6. С. 67-77. 4. *Катренко С.А., Буров Є.В.* Застосування формальних моделей та методів у видобуванні та виявленні знань зі сховищ даних // "Інформаційні системи та мережі": Вісник Національного університету "Львівська політехніка". 2000. №406. С. 156-163. 5. *Han J., Kamber M.* Data mining: concepts and techniques. Morgan Kaufman Publishers, 2000. 6. *Месарович М., Такахара Я.* Общая теория систем: математические основы. М.: Мир, 1978. 312 с. 7. *Мейер Д.* Теория реляционных баз данях: Пер. с англ. М.: Мир, 1987. 608 с. 8. *Кравець Р.Б.* Багатовимірна модель даних у системах аналітичної обробки інформації. // "Інформаційні системи та мережі": Вісник Національного університету "Львівська політехніка". 1998. №330. С. 147-153.

Поступила до редколегії 18.05.2001