

УДК 004.82

ПРИНЦИПИ ФОРМУВАННЯ БАЗИ ЗНАНЬ

С.В. Голуб, І.А. Жирякова

Черкаський національний університет ім. Б. Хмельницького

e-mail: irena_zh@ukr.net

Інженерія знань – досить молодий напрям штучного інтелекту, який з'явився внаслідок досліджень, пов'язаних з практичною розробкою інтелектуальних систем, основним елементом яких є база знань. Реалізація бази знань передбачає використання єдиної концепції, яка описувала б способи формалізації та роботи зі знаннями. Адже, знання є основою будь-якої інтелектуальної системи та являють собою добре структуровані дані або дані про дані. Але незважаючи на те, що питанням роботи зі знаннями та застосуванню баз знань у різних предметних галузях присвячено досить багато публікацій (з яких оглядовими є [1-7]), у науковому співтоваристві досі немає згоди щодо раціональності застосування тієї чи іншої моделі представлення знань для розв'язання практичних задач, які зводяться до наступних класів: продукційні моделі, семантичні мережі, фреймові моделі і формальні логічні моделі.

Базою знань зазвичай називають сукупність концептів і зв'язків між ними. Одна з проблем баз знань полягає в їх неповноті. Таким чином, виникає необхідність автоматичного наповнення бази знань для усунення неповноти за рахунок додавання в неї нових концептів. Сучасні методи вирішення даної проблеми припускають наступний загальний алгоритм наповнення бази знань:

- Отримання нових концептів з колекції певної предметної області.
- Визначення чи є в базі знань їх аналоги і, якщо вони не знайдені, то перейти до наступного кроку; якщо ж знайдені – то перейти до попереднього кроку.
- Визначити зв'язки між концептами.
- Обчислити формальні аксіомати для концептів.

Розпізнавання нових та існуючих концептів може відбуватися по-різному, залежно від масштабу бази знань. У разі великих баз знань, що покривають різні предметні області, можуть виникати ситуації, коли для частини нових концептів аналоги вже існують в базі. Тому виникає необхідність у більш складних процедурах розпізнавання концептів, а власне і формуванні нових принципів роботи з ними.

Отже, в задачу даного дослідження входить розробка нового підходу до формування баз знань. Даний підхід припускає інші принципи збереження знань, а тому дозволить спростити процедуру отримання нових знань.

Процес роботи з даними в будь-якій інтелектуальній системі зазвичай проходить в чотири етапи: відбір даних, їх трансформація, аналіз та вибір стратегії використання (рис. 1).

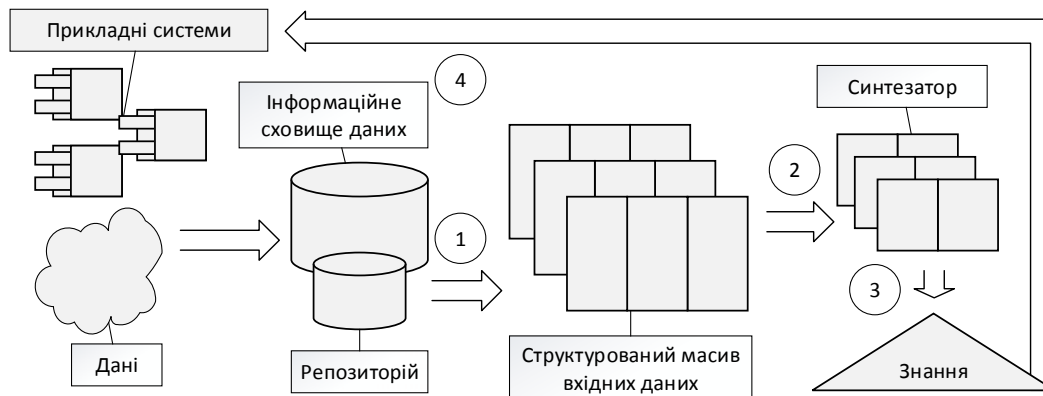


Рис. 1. Архітектура системи інтелектуального аналізу даних, де 1-4 етапи аналізу

Розглянемо докладніше призначення кожного етапу.

Відбір даних (1). Для вирішення поставленої задачі зазвичай потрібні не всі дані, які надходять ззовні та розміщуються в інформаційному сховищі. Отже, спочатку необхідно вибрати підмножину, яка буде підлягати аналізу. При цьому, можливо, буде потрібно об'єднати декілька таблиць, а отримані записи потім відфільтрувати.

На даному етапі можна скористатись відомими алгоритмами кластеризації. Адже, кластеризація – це процес розбиття даних на ряд сегментів, або кластерів, які об'єднують лише ті дані, що мають загальні характеристики чи властивості. Результати кластеризації застосовуються для підведення сумарних підсумків по кластерам і в якості вхідних даних для будь-якого іншого методу аналізу. Оскільки розміри кластерів менше розмірів інформаційного сховища даних і в них містяться відносно «чисті» дані, час їх обробки помітно скорочується, а якість моделей побудованих на таких даних підвищується.

Трансформація даних (2). Після підготовки робочих таблиць зазвичай проводиться попередня обробка даних, характер якої визначається методами, що застосовуються в ході аналізу. Трансформація може полягати у видаленні зашумлених даних і дублюючих записів, перетворенні типів даних, додаванні нових атрибутів тощо.

Аналіз (3). На основі трансформованих даних синтезуються індуктивні моделі, які ієрархічно поєднуються та послідовно обробляються одним або декількома методами з метою отримання необхідної інформації або нових знань. Отже, історична інформація у вигляді структурованого масиву вхідних даних, що надходить з репозиторія у вигляді часових рядів, відображає динаміку досліджуваної системи в минулому. І якщо вдається побудувати математичну модель, яка адекватно описує цю динаміку, є ймовірність, що з її допомогою можна передбачити поведінку системи в майбутньому. Крім того, цінність прогнозних моделей, які базуються на індукційних методах, полягає не тільки в простоті представлення і тлумачення отриманих знань, а й у їх досить високій точності.

Стратегії використання (4). Отримані на етапі аналізу знання будуть обиратись для подальшого використання в ході розв'язання поставлених задач з підключенням експерта-аналітика.

Концептуалізація знань як засіб формалізації, повинна сприяти висхідному синтезу ієрархічних структур (сукупності аналітичних моделей, які б давали можливість індуктивного моделювання нових знань) елементами яких були б відповідні поля знань. Отже, до основних етапів даного процесу (рис. 2) можна віднести: формування ієрархічних структур, тестування та їх оперативне використання в процесі прийняття рішень.

Згідно запропонованого підходу кожен масив вхідних даних (МВД 1 – МВД N) (1), який відповідає певному виду багатовимірних даних, отриманих в процесі дослідження обирається у відповідності до поставленої задачі з бази даних (БД).

$$D = \{(x_i, y_i)\}_{i=1}^N, x \in R^t \quad (1)$$

Далі кожен МВД аналізується синтезатором (С) на предмет встановлення взаємозв'язків та побудови алгоритму синтезу моделі (АСМ 1 – АСМ N), які в комплексі формують поле знань на кожному рівні ієрархії. Отримані ієрархічні структури, являють собою фрагмент знань (образ) орієнтований на доступ до структурованої на попередньому етапі інформації, представленої в БД.

В результаті, на кожному рівні відбувається індуктивний висхідний синтез, який породжує фрагмент знань, що являє собою дедалі складнішу комбінацію. Контролер (К) забезпечує точність і адекватність кожного фрагменту оптимальної складності.

Складність комбінацій на кожному рівні обробки інформації зростає в залежності від кількості нових образів, які враховуються АСМ. На верхньому рівні ієрархії формується рішення поставленої задачі та вагові коефіцієнти (частинні похідні моделей по кожній змінній), які вносяться до сховища даних (СД).

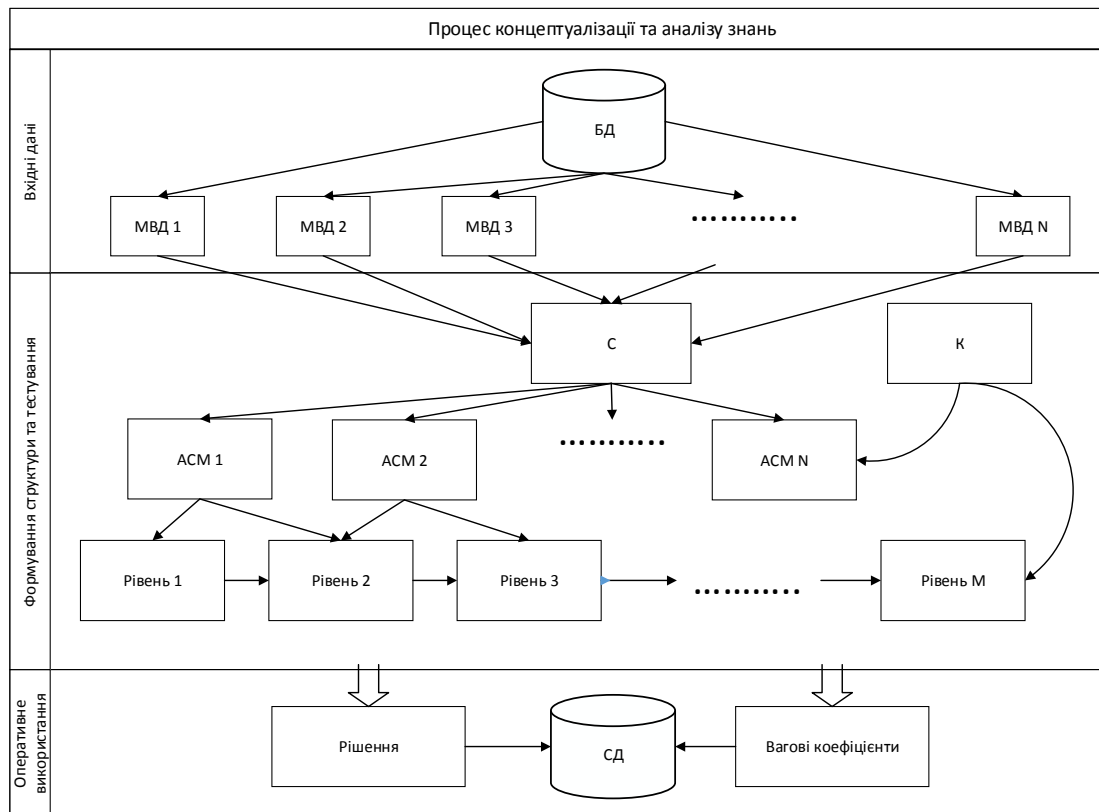


Рис. 2. Концептуалізація та аналіз знань

Список літератури

1. Knowledge Acquisition Tools, Methods, and Mediating Representations [Text] / [Motoda H., Mizoguchi R., Boose J.H., Gaines B.R.]. – Proceedings of the First Japanese Knowledge Acquisition for Knowledge-Based Systems Workshop: JKAWS-90, Ohmsha Ltd: Japan, 1990. – 32 p.
2. Wielinga B. KADS: A Modeling Approach to Knowledge Engineering [Text] / B. Wielinga, G. Schreiber, J.A. Breuker // In Knowledge Acquisition. – 1992. – vol. 4, n. 1. – P. 5-53.
3. Tuthill G.S. Knowledge Engineering [Text] / G.S. Tuthill. – TAB Books Inc., 1994. – 750 p.
4. Adeli H. Knowledge Engineering [Text] / H. Adeli. – New-York: McGraw-Hill Publishing Company, 1994. – 914 p.
5. Приобретение и формализация знаний [Текст] / [Аверкин А.Н., Блишун А.Ф., Гаврилова Т.А., Осипов Г.С.] / Искусственный интеллект: [в 3 т.] / под ред. Д.А. Поспелова. – М.: Радио и связь, 1990. – Т. 2: Модели и методы. – 304 с.
6. Krishna S. Introduction to Database and Knowledge-base Systems [Text] / Krishna S. – Singapore ; River Edge, NJ : World Scientific Series in Computer Science, 1992. – 972 p.
7. Stefik M.J. Introduction to Knowledge Systems [Text] / Stefik M.J. – San Francisco, CA: Morgan Kaufmann, 1995. – 896 p.
8. Erk K. Unknown word sense detection as outlier detection // Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006. – P. 128–135.