Математичне моделювання електротехнічних систем

Лекція 4

3.1 Data Mining – інтелектуальний аналіз даних

У зв'язку з вдосконаленням технологій запису і зберігання даних на людей обрушилися колосальні інформаційні потоки в самих різних областях.

Стало ясно, що без продуктивної переробки потоки сирих даних утворюють

нікому не потрібне звалище.

Специфіка сучасних вимог до переробки інформації наступна:

- Дані мають необмежений об'єм;
- Дані є різнорідними (кількісними, якісними, текстовими);
- Результати повинні бути конкретні і зрозумілі;
- Інструменти для обробки сирих даних повинні бути прості у використанні.

Традиційна математична статистика, не справляється повною мірою із завданням переробки інформації. Головна причина — концепція усереднювання по вибірці, що приводить до операцій над фіктивними величинами (типу середньої температури пацієнтів по лікарні, середньої висоти будинку на вулиці і тому подібне).

Методи математичної статистики виявилися корисними головним чином для перевірки заздалегідь сформульованих гіпотез (verification-driven data mining) і для “грубого” розвідувального аналізу, що становить основу оперативної аналітичної обробки даних (online analytical processing, OLAP).

Термін Data Mining отримав свою назву з двох понять: пошуку цінної інформації у великій базі даних (data) і здобичі гірської руди (mining).

Термін Data Mining часто переводиться як добування даних, витягання інформації, розкопка даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, витягання знань, аналіз шаблонів. Поняття "Виявлення знань в базах даних" (Knowledge Discovery in Databases, KDD) можна вважати синонімом Data Mining.

Методи Data Mining (або, що те ж саме, Knowledge Discovery In Data, скорочено, KDD) лежать на стику баз даних, статистики і штучного інтелекту.

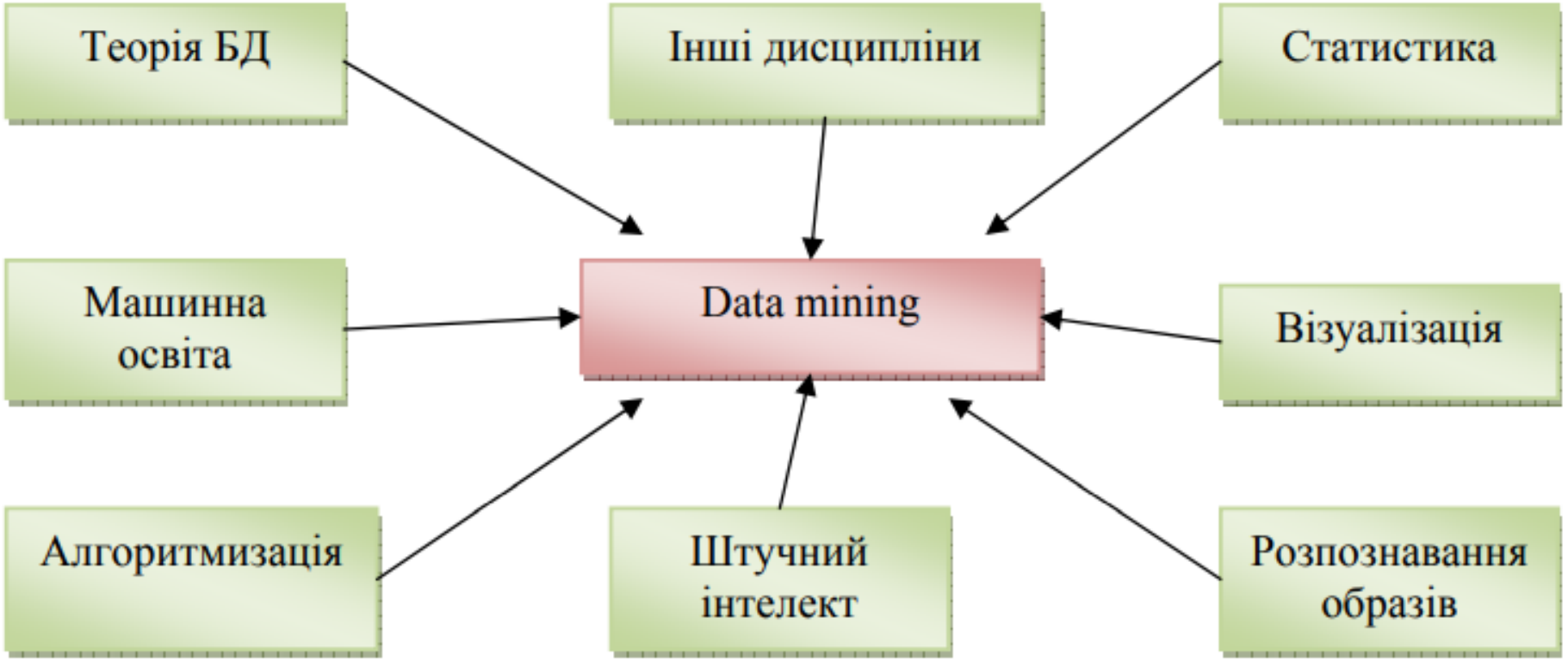

Рисунок 3.1. – Data mining як мультидисциплінарна область

Приведемо короткий опис деяких дисциплін, на стику яких з'явилася технологія Data Mining.

Поняття Статистики

Статистика - це наука про методи збору даних, їх обробки і аналізу для виявлення закономірностей, властивих явищу, що вивчається.

Статистика є сукупністю методів планування експерименту, збору даних, їх уявлення і узагальнення, а також аналізу і отримання виводів на підставі цихданих.

Статистика оперує даними, отриманими в результаті спостережень або експериментів.

Поняття Машинного навчання

Єдиного визначення машинного навчання на сьогоднішній день немає. Машинне навчання можна охарактеризувати як процес отримання програмою нових знань. Мітчелл в 1996 році дав таке визначення: "Машинне навчання - це наука, яка вивчає комп'ютерні алгоритми, що автоматично поліпшуються під час роботи". Одним з найбільш популярних прикладів алгоритму машинного навчання є нейронні мережі.

Поняття Штучного інтелекту

Штучний інтелект - науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.

Термін інтелект (intelligence) походить від латинського intellectus, що означає розум, розумові здібності людини.

Відповідно, штучний інтелект (AI, Artificial Intelligence) тлумачиться як властивість автоматичних систем брати на себе окремі функції інтелекту людини. Штучним інтелектом називають властивість інтелектуальних систем виконувати творчі функції, які традиційно вважаються прерогативою людини. Кожен з напрямів, Data Mining, що сформували, має свої особливості. Проведемо порівняння з деякими з них

Порівняння статистики, машинного навчання і Data Mining .

Рисунок 3.2. – Порівняння статистики, машинного навчання і Data Mining

Поняття Data Mining близько пов'язане з технологіями баз даних.

Поняття Data Mining

Data Mining - це процес підтримки ухвалення рішень, заснований на пошуку в даних прихованих закономірностей (шаблонів інформації).

Технологію Data Mining достатньо точно визначає Григорій Патецький Шапіро (Gregory Piatetsky-Shapiro) – один із засновників цього напряму:

Data Mining- це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності. Приведемо ще декілька визначень поняття Data Mining. Data Mining- це процес виділення з даних неявної і неструктурованої інформації і представлення її у вигляді, придатному для використання.

Data Mining- це процес виділення, дослідження і моделювання великих об'ємів даних для виявлення невідомих до цього структур (patterns) з метою досягнення переваг в бізнесі (визначення SAS Institute).

Data Mining- це процес, мета якого – виявити нові значущі кореляції, зразки і тенденції в результаті просіювання великого об'єму даних, що зберігаються, з використанням методик розпізнавання зразків плюс застосування статистичних і математичних методів (визначення Gartner Group).

Суть і мету технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих об'ємах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Неочевидних - це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

Об'єктивних - це означає, що виявлені закономірності повністю відповідатимуть дійсності, на відміну від експертної думки, яка завжди є суб'єктивною.

Практично корисних - це означає, що виводи мають конкретне значення, якому можна знайти практичне застосування.

Знання - сукупність відомостей, яка утворює цілісний опис, відповідний деякому рівню обізнаності про описуване питання, предмет, проблемі і так далі

Використання знань (knowledge deployment) означає дійсне застосування знайдених знань для досягнення конкретних переваг

У основу технології Data Mining покладена концепція шаблонів (patterns), які є закономірностями, властивими під вибіркам даних, які можуть бути виражені у формі, зрозумілій людині.

"Mining" по-англійськи означає "видобуток корисних копалин", а пошук закономірностей у величезній кількості даних дійсно схожий на цей процес.

Мета пошуку закономірностей – представлення даних у вигляді, що відображає шукані процеси. Побудова моделей прогнозування також є метою пошуку закономірностей

Важливе положення Data Mining — не тривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відображати неочевидні, несподівані (unexpected) регулярності в даних, такі, що становлять так звані приховані знання (hidden knowledge).

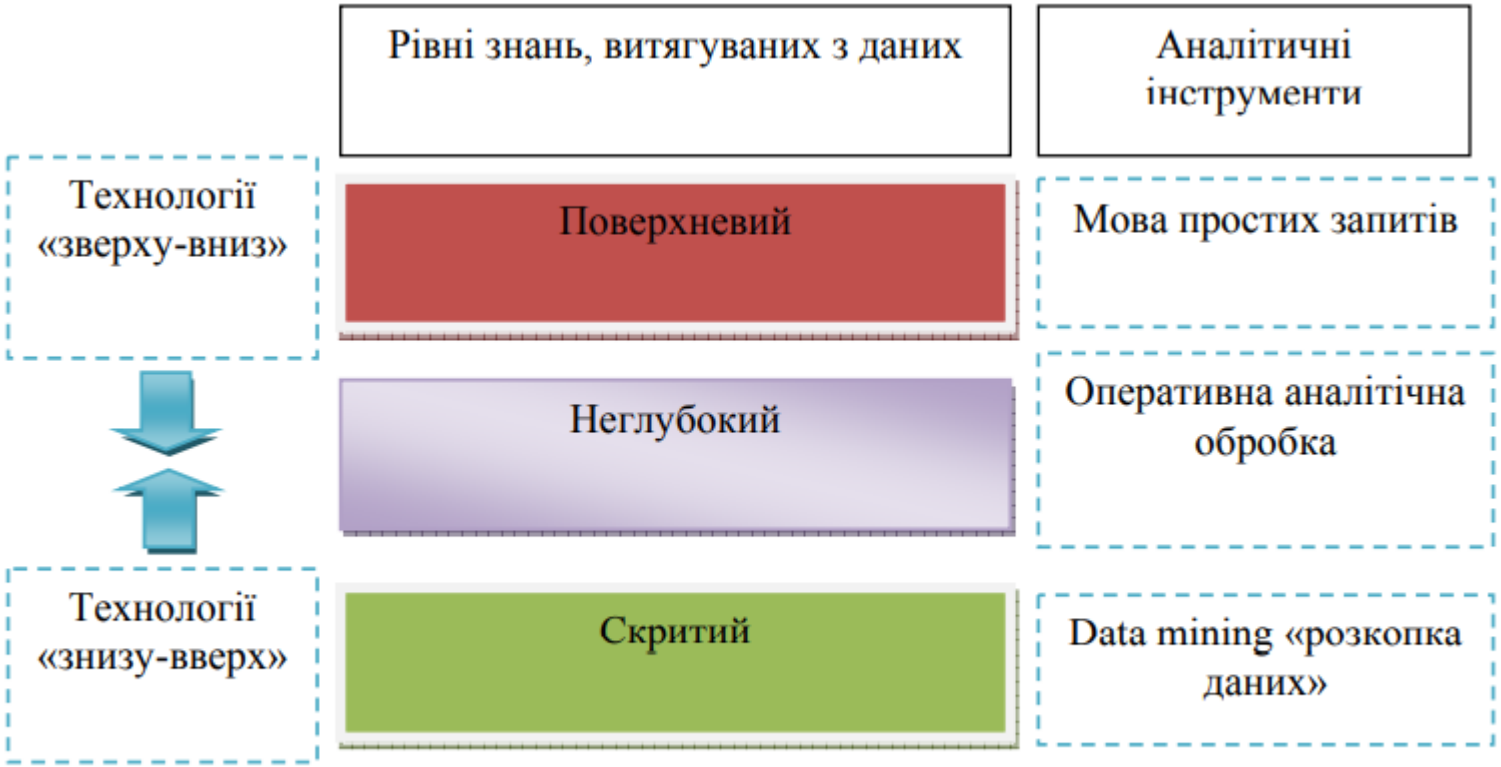

Рисунок 3.3 – Рівні знань, витягуваних з даних

Історично склалося, що у терміну Data Mining є декілька варіантів перекладу (і значень):

- витягання, збір даних, здобич даних (ще використовують Information Retrieval або IR);
- витягання знань, інтелектуальний аналіз даних (Knowledge Data Discovery або KDD, Business Intelligence).

Найчастіше витягання даних (збір) є підготовчим етапом для витягання знань (аналіз).

Завдання, вирішувані Data Mining:

- Класифікація — віднесення вхідного вектора (об'єкту, події, спостереження) до одного із заздалегідь відомих класів.

- Кластеризація — розділення безлічі вхідних векторів на групи (кластери) по ступеню «схожості» один на одного.

- Скорочення опису — для візуалізації даних, спрощення рахунку і інтерпретації, стиснення об'ємів збираної інформації, що зберігається.

- Асоціація — пошук зразків, що повторюються. Наприклад, пошук «стійких зв'язків в корзині покупця».

- **Прогнозування** – знаходження майбутніх станів об'єкту на підставі попередніх станів (історичних даних).

- **Аналіз відхилень** — наприклад, виявлення нетипової мережевої активності дозволяє виявити шкідливі програми.

- **Візуалізація даних.**

Information retrieval

Information retrieval використовується для отримання структурованих даних або репрезентативної вибірки меншого розміру. Information retrieval оперує даними першого рівня, а в результаті видає інформацію другого рівня.

Найпростішим прикладом information retrieval є пошукова система, яка на підставі якихось алгоритмів виводить частину інформації з повного набору документів. Крім того, будь-яка система, яка працює з тестовими даними, метаінформацій або базами даних тим або іншим способом використовує інструменти information retrieval. Інструментами можуть виступати методи індексації, фільтрації, сортування даних і так далі.

Text Mining

Інші назви: text data mining, text analysis, дуже близьке поняття – concern mining.

Text mining може працювати як з сирими даними, так і з частково обробленими, але на відміну від information retrieval, text mining аналізує текстову інформацію за допомогою математичних методів, що дозволяє отримувати результат з елементами знання.

Завдання, які вирішує text mining: знаходження шаблонів даних, отримання структурованої інформації, побудова ієрархій об'єктів, класифікація і кластеризація даних, визначення тематики або області знань, автоматичне реферування документів, завдання автоматичної фільтрації контента, визначення семантичних зв'язків та інші.

Для вирішення завдань text mining використовують статистичні методи, методи інтерполяції, апроксимації і екстраполяції, нечіткі методи, методи контент-анализа.

Web Mining

web mining – набір підходів і техніки для витягання даних з веб-ресурсів. Оскільки веб-джерела, як правило, не є текстовими даними, то і підходи до процесу витягання даних відрізняються в цьому випадку. Інформація у WEB зберігається у вигляді спеціальної мови розмітки HTML (хоча є і інші формати – RSS, Atom, SOAP), веб-сторінки можуть мати додаткову метаінформацію, а також інформацію про структуру (семантиці) документа, кожен веб-сервер документ знаходиться усередині якогось домена і до нього можуть застосовуватися правила пошукової оптимізації (SEO).

Для вирішення вищеописаних завдань використовуються різні методи і алгоритми Data Mining. З огляду на те, що Data Mining розвивалася і розвивається на стику таких дисциплін, як статистика, теорія інформації, машинне навчання, теорія баз даних. Більшість алгоритмів і методів Data Mining були розроблені на основі різних методів з цих дисциплін. Наприклад, процедура кластеризації k-means була просто запозичена із статистики. Велику популярність отримали наступні методи Data Mining: нейронні мережі, дерева рішень, алгоритми кластеризації, у тому числі і масштабовані, алгоритми виявлення асоціативних зв'язків між подіями і так далі.

3.2 Історичний екскурс Data Mining