УДК 004.04

Лисак Володимир Васильович,
Київської державної академії водного транспорту ім. гетьмана П. Конашевича – Сагайдачного

ІНЖЕНЕРІЯ АНАЛІТИЧНИХ ЗНАНЬ

Розглянуто еволюцію аналітичних систем. Запропоновано поняття аналітичних знань як результатів методів обробки даних, а такожінженерний підхід до їх отримання з баз даних завдяки представленню всіхметодів обробки у виглядіграфічних об'єктів. Об'єкти мають клієнт-серверну архітектуру, інкапсулюють в собі вхідні дані, методи обробки даних з обраними режимами обробки, а також результати у вигляді таблиць, графіків та схем. Серед об'єктів представленіприклади методів підготовки даних (вибірки з різних джерел, узгодження, перевірка на коректність, вирізування даних з таблиць, транспонування тощо), а також методів пошуку знань та залежностей. Будучі одночасно і клієнтом і сервером аналітичні знання можуть створювати ланцюжки виконання методів. Послідовне застосування методів створило можливість концентрувати аналітичну інформацію і отримувати комбіновані аналітичні знання.

Рассмотрена эволюция аналитических систем. Предложено понятие аналитических знаний как результатов методов обработки данных, а также инженерный подход их получения из баз данных благодаря представлению всех методов обработки в виде графических объектов. Объекты имеют клиент-серверную архитектуру, инкапсулируют в себе входные данные, методы обработки данных с выбранными режимами обработки, а также результаты в виде таблиц, графиков и схем. Среди объектов представлены примеры методов подготовки данных (выборки из различных источников, согласования, проверка на корректность, выборки данных из таблиц, транспонирование и т.д.), а также методов поиска знаний и зависимостей. Будучи одновременно и клиентом и сервером аналитические знания могут создавать цепочки выполнения методов. Последовательное применение методов создало возможность концентрировать аналитическую информацию и получать комбинированные аналитические знания.

The evolution ofanalytical systemshas been considered.The concept of analytical knowledge as a result of data processing and engineering approach to their receipt of databases through presentation of all methods of treatment in the form of graphic object. Objects have a client-server architecture that encapsulate a data input, data processing with selected modes of processing, and the results in the form of tables, graphs and diagrams. Among the objects to be methods of training data (samples from different sources, alignment, checking for correctness, cutting data tables, transpose, etc.), as well as methods of seeking knowledge and relationships. Being both a client and server analytical knowledge can create chains of execution methods. Consistent application of methods created to concentrate analytical information and receive a combined analytical knowledge.

Ключові слова: аналітичні системи, інженерія знань, методи пошуку знань

Вступ

Сучасне суспільство все більше називають суспільством, що базується на знаннях, тим самим підкреслюючи важливу роль знань, науки та інформації в суспільному житті. Для сьогодення характерна глобалізація змін в економіці, науці, освіті завдяки поширенню нових технологій здійснення та пришвидчення комунікацій. Сьогодні вже процеси виробництва та інженерії знань не виступають окремо, а стають необхідною фундаментальною умовою функціонування та розвитку суспільства.

В той же час сучасний діловий світ розуміє, що без глибокого аналізу інформації, яка переповнює ринки та внутрішню діяльність організацій, неможливе успішне ведення бізнесу. Потоки інформації при її кваліфікованій обробці, аналізі та синтезу висновків, здатні надати підприємству конкурентні переваги по відношенню до інших учасників ринку, які її не мають.

Актуальність розвитку аналітичних комп'ютерних систем підтверджується тенденціями напрямків розробок у світових лідерів програмного забезпечення, а саме: на ринки програмного забезпечення виходять інформаційно-аналітичні системи для використання у різних сферах. Аналітичні системи – BI (Business Intelligence[1]) – представлені цілою лінійкою рішень більш ніж 10 різних компаній. За думкою експертів IDC[2],

[1] Business intelligence або скорочено BI — бізнес-аналіз, бізнес-аналітика. Під цим поняттям частіше всього розуміють програмне забезпечення, призначене для допомоги менеджеру в аналізі інформації про свою компанію.

[2] IDC - International Data Corporation є провідним світовим постачальником ринкової

еволюцію інформаційно-аналітичних систем можна розділити на 3 хвилі:

1 – (до 1990 року) збір інформації та підготовка звітності;

2 – (1990-2005 рік) розвиток швидкого багатомірного аналізу на базі технології OLAP [1], а також самостійне створення нерегламентованої звітності;

3 – (з 2005 року) створюється акцент на розвиток прикладних засобів застосування, що включає аналіз, прогностику та пошук прихованої інформації (методи Data Mining [2]).

Сьогодні ми як раз бачимо появу BI-решень третьої хвилі. До них можна, наприклад, віднести BI-системи від відомих світових лідерів: Oracle, IBM, Microsoft, SAS тощо.

Що стосується перспектив використання технологій, то зараз пост-радянський ринок як раз стоїть на порозі початку впровадження BI-систем третьої хвилі, які здатні шукати приховану інформацію, будувати прогнозну аналітику та проводити перехресний аналіз інформації з несумісних на перший погляд джерел даних. Самі експерти IDC вважають, что ці хвилі мають 15-річний цикл, але, приймаючи до уваги швидкість, з якою на світовому ринку з'являються нові решення, можна очікувати, что третя хвиля все ж закінчиться раніше і вже через пару років пост-радянський ринок стане свідком появи рішень нового, четвертого покоління. Можна очікувати, що це буде більш тісна інтеграція систем класу підтримки прийняття рішень DSS (Decision Support System), систем прогнозування та пошуку прихованих залежностей (Data Mining). Іншими словами, користувачу буде надана можливість вибору сценарія розвитку ситуації, виходячи з якого система сама проведе аналіз накопиченої інформації, побудує прогноз зміни ключових показників та запропонує оптимальні варіанти дій, які б привели до ліпшого результату. Тобто BI-система позбавить користувача від необхідності виконання довгої рутинної роботи з пошуку причинно-наслідкових зв'язків при аналізі даних, передачі результатів роботи однієї системи в іншу, контролі коректності завантаженої інформації тощо. Ці завдання будуть виконуватися автоматично - від користувача буде потрібно всього лише на початку роботи вибрати сценарій розвитку тієї чи іншої задачі, а в кінці - найбільш вподобаний йому оптимальний варіант вирішення цього завдання, що відповідає змісту фрази "Business Intelligence".

Інформаційно-аналітичні системи являють собою надбудову над тими інформаційними системами, що вже функціонують на підприємствах, вони не впливають на їх функціонування та не вимагають їх заміни. Ключовою функцією цих систем є обробка, акумулювання та аналіз даних з усіх видів діяльності організації чи окремої системи, отримання закономірностей, залежностей, прогнозування та оцінці можливих ризиків, оптимізації діяльності.

Фактично функціями таких систем є з одного боку - робота з даними (пошук, формування вибірок з таблиць, їх склеювання, розрізування, транспонування, створення зведених та агрегованих таблиць), а з іншого - їх перетворення до вигляду, з якого можна отримати нові знання в тематичній області після застосування різних математичних методів або практичних алгоритмів, які застосовуються для прийняття рішень або системного моделювання процесів чи явищ.

Метою цієї статті є опис та реалізація інженерного підходу до створення аналітичних знань, відокремлення інженера знань від експерта у тематичній області та математика з його поглибленими знаннями у математичних методах обробки даних. Таким чином робота з даннними для отримання аналітичної інформації (знань) повинна стандартизуватись та формуватись як інженерна дисципліна з основним виконавцем – інженером знань.

Уточнення поняття «аналітичні знання»

З розвитком структур виникла концепція *знань*, у відповідності до якої об'єднуються блоки різних типів інформації, а перехід від даних до знань - логичний наслідок розвитку та вдосконалення інформаційно - логічних структур, що оброблюються на ЕОМ [3]. Що ж таке «знання»? Поняття «знання» різними авторами трактується по-різному, але як таке, що найбільше відповідає тематиці статті, то це - виявлені закономірності в предметній області (принципи, зв'язки, закони), що дозволяють розв'язувати тематичні задачі. Поняття

інформації, консультаційних послуг та організатор заходів на ринках інформаційних технологій, телекоммунікацій та користувацької техніки. Більш ніж 1000 аналітиків IDC, глобальний, регіональний та місцевий досвід в области технології в більш ніж 110 країнах світу.

«знання» прийнято поділяти на декларативні та процедурні. Декларативні знання - це певна множина тверджень, незалежно від того де і коли вона використовується. **Процедурні знання** або **правила** являють собою набір певних процедур перетворення даних в знання. Поділ знань на декларативні та процедурні достатньо умовний і відомі моделі представлення знань використовують у різній мірі ті або інші поняття.

Формування знань виконується із залученням експерта – фахівця в тематичній області. А задача інженера - це виявити каркас поглядів та висновків експерта. Інженер із знань може спиратися на дві найбільш популярні теорії мислення – логічну та асоціативну. Якщо логічна, як правило, використовується в розробках з штучного інтелекту, то асоціативна будується з минулого досвіду, методом спроб та помилок, а також методом аналогій. І тут можна говорити про аналітичні *знання,* які принципово відрізняються від інших перш за все способом їх отримання. Як назвати результати застосування однакових процедурних знань до різних даних? Звичайно отримані результати будуть різні, і вони будуть претендувати на назву *аналітичні знання*. Будемо вважати, що аналітичні знання - це знання, які отримані як результат застосування процедурних знань, а саме: методів Data Mining[3] (розв'язок задачі пошуку асоціацій, послідовностей, задачі класифікації, кластерізації, задач прогнозування та оптимізації) або методів багатомірної статистики, що застосовуються до статистичних даних в базах даних, або певних моделей, наприклад, нейромережевих моделей чи генетичних алгоритмів тощо.

Системи аналізу та обробки наукової інформації, або системи отримання аналітичних знань будуються на принципах та технологіях, відмінних від принципів побудови інших інформаційних систем сучасного підприємства, тому що дані, які використовуються, вимагають проведення спеціальної їх підготовки, надання їм коректності для подальшого використання.

Постановка проблеми.

Користувачі всередині організацій потребують різних типів інтерфейсів в залежності від типу інформації, з якою вони працюють. Керівники потребують спеціальної інформації, що подається за допомогою інтуїтивно зрозумілого інтерфейсу, з точними показниками досліджуваних експериментальних даних. Бізнес-аналітикам необхідна можливість працювати безпосередньо з даними, використовуючи інтерфейс орієнтований на певну задачу.

Необхідно створити таке універсальне середовище - базу *аналітичних знань*, в якій можна виконувати елементарну обробку даних: використовувати вибірки даних для аналізу з різних, можливо зовнішніх, сховищ, узгоджувати їх для подальшої обробки, виявляти та автоматично коригувати їх структуру, застосовувати методи отримання знань та зберігати у вигляді аналітичних знань. І все це завдяки графічному інтуїтивно зрозумілому інтерфейсу, який буде зручним як для керівника організації, бізнес-аналітика, так і для інженера знань.

Необхідно також щоб аналітична система не вимагала спеціальних математичних знань, але дозволяла галузевому аналітику або інженеру знань виконувати пошук знань та залежностей за допомогою новітніх методів. Окрім того для вдосконалення отримання аналітичних знань необхідно, щоб без втручання «програміста» можна було б в одному результаті послідовно акумулювати результати послідовної роботи окремих методів обробки.

Основні принципи побудови аналітичної системи

Для того, щоб система управління данними претендувала на назву аналітичної, необхідно її спроектувати як базу «знань» з різними методами обробки даних та пошуку знань.

Отже, якою повинна бути базова одиниця інформації в базі аналітичних знань? Легше за все можна вивчити роботу аналітиків (інженерів знань), де виконується кропітка робота з підготовкою даних (вибірки з різних джерел, узгодження, перевірка на коректність, вирізування даних з таблиць, транспонування тощо) – всі ці методи обробки повинні бути присутні в аналітичній базі знань і використовуватись як універсальні аналітичні об'єкти на рівні з процедурними знаннями. В той же час, якщо взяти будь-який аналітичний звіт – в ньому присутні результати математичних експериментів, статистичних

3 Одне з найважливіших призначень методів Data Mining полягає в наочному поданні результатів обчислень, що дозволяє використовувати інструментарій Data Mining людьми, які не мають спеціальної математичної підготовки.

методів оцінки окремих змінних, їх порівнянь, знаходження залежністей, групування або класифікація за різними ознаками, які зменшують кількість змінних створюючи більш складні структури. Застосовуються також різні моделі, методи прогнозування та оптимізації.

Таким чином основним елементом такої бази знань повинен бути ***графічний об'єкт***, який об`єднує метод обробки чи алгоритм розрахунку з можливістю динамічного підключення вхідних даних з різних джерел та збереження результатів застосування методу обробки як у вигляді таблиці, так і у вигляді схеми або графіка, в зрозумілій, легко прийнятній аналітиком формі(дивись рис.1). Такий об'єкт будемо називати «аналітичним знанням» або просто «знанням».

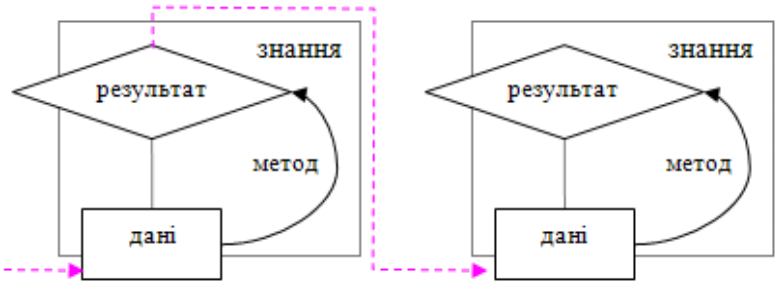

Рис.1. Схема базових елементів «знання», які включені в послідовний ланцюжок розрахунків

Складовими елементами «знання» є:

- дані – вхідна інформація, яка необхідна для виконання певного методу розрахунку;
- метод (розрахунку) – деякий статистичний, аналітичний або математичний метод, який є серед множини існуючих, а також той, який завжди можна додати до використання; це може бути також метод обробки та підготовки даних;
- результат – інформація або дані, які отримані шляхом застосування методу обробки чи розрахунку до вхідних даних. Результат може слугувати вхідною інформацією для іншого знання, забезпечуючи послідовну обробку різних методів.

Кожний об'єкт - «знання» забезпечено стартером для автоматизованого запуску на виконання поточного об'єкту в залежності від сигналу «виконано» попередніх об'єктів. Таким чином забезпечується потокова обробка «знань», яка формується автоматично відповідно до створення сценарію послідовної обробки даних інженером знань (дивись рис.2)

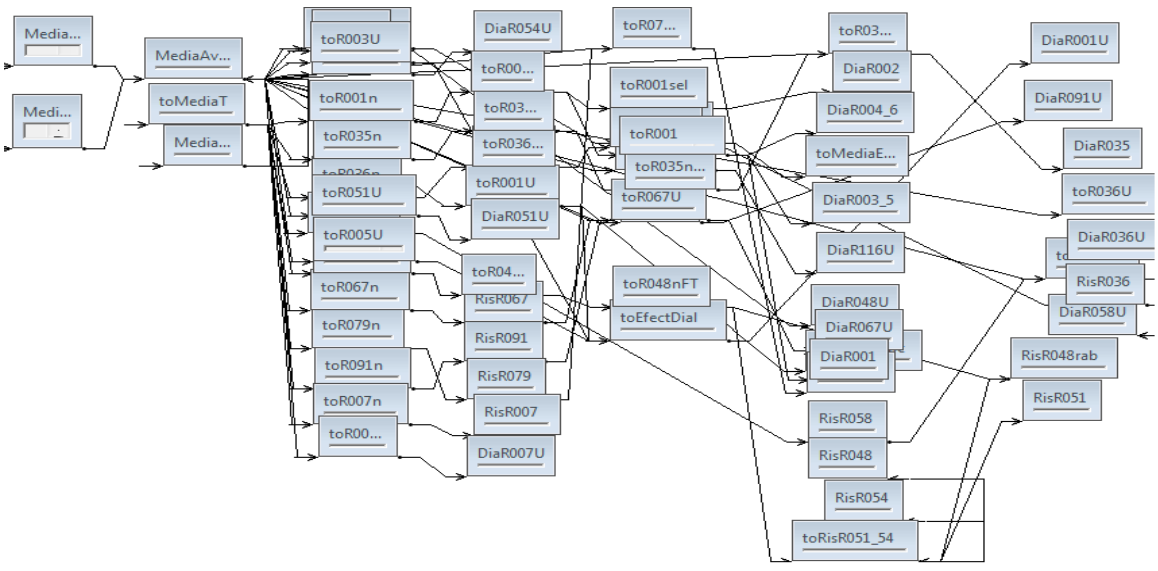

Рис.2. Схема потокової обробки «знань», яка формується автоматично

Розвиток базового об’єкту в майбутньому не торкнеться його суті, але можливо його розширити і наділити декількома результатами, а також поєднати декілька методів послідовно для утворення нового методу - ланцюжка методів.

Інженерія аналітичних знань повинна включати стандартні підходи до обробки даних, зокрема, якщо виконується аналіз невідомих данних.

Системний підхід до аналізу даних вимагає, щоб з одного боку об’єктами в базі були елементарні дії роботи над перетворенням таблиць, а з другого - методи пошуку знань.

Засоби попередньої обробки даних, які для інженера знань забезпечуть всі можливі операції з даними, і є відносними, це:

- запити на вибірку із зовнішніх таблиць, об’єднання таблиць, склеювання, транспонування, створення зведених таблиць, чи навпаки розведених (з прямокутних таблиць створюються подовжені), при чому дії без явного вказування назв полів;

Після підготовки даних до них можна застосовувати системний аналіз даних [4], різноманітні методи аналізу та перетворення даних в знання, саме:

- дослідження структур даних (основні статистики, групові середні, частотний розподіл, таблиці перерізів), їх порівняння;
- виявлення статистично-значущих зв’язків між змінними, методи асоціацій та послідовностей (кореляційний аналіз, факторний аналіз, парна регресія);
- групування, класифікація та кластерізація (кластерний аналіз, дерево рішень, нейромережі);
- знаходження факторів для зменшення змінних (факторний аналіз);
- визначається поведінка окремих груп, прогнозування змін, порівняльний аналіз (екстраполяція, динамічні ряди, нейромережі, множинна регресія);
- методи оптимізації (метод Тагучі та нейромережі);

Якщо порівнювати інженера знань, який створює ланцюжки обробки даних та отримання знань з осмисленою роботою науковця, то необхідно аналітичну базу знань забезпечити можливістю оформлення в один об’єкт ланцюжка методів, що повторюється. Таким чином в системі підтримується концепція «виникнення» нових складних структур обробки даних. Це також вписується в поняття того, що будь-який розвиток припускає споживання найпростіших елементів для поступової трансформації і переходу на більш високий рівень[5].

Результати.

Таким чином було розроблене ядро системи з потоковою обробкою даних, в якій база аналітичних знань являє собою набір засобів доступу до зовнішніх стандартних баз даних, забезпечує створення вибірок, засобів обробки даних, методів аналітичної обробки. Одиницею такої бази є складний об’єкт, який інкапсулює в собі вхідні дані, методи обробки та їх режими, а також результат обробки. При цьому результат представляється не тільки у вигляді розрахунку, а і у візуальній формі – у вигляді графіка, діаграми чи схеми. В ній поряд із поширеними стандартними методами обробки даних, методів математичної статистики та пошуку знань, використовуються розроблені нові аналітичні моделі, як результат послідовного розрахунку декількох методів. Система має графічний інтерфейс, зручний для користувачів різного професійного рівня.

Приклад найпростішої потокової обробки даних, або інженерії аналітичних знань:

- два вхідних об”єкта – таблиця кодифікатор та результат розрахунку середніх значень (знання «Основні Статистики»)
- після їх поєднання - результат (знання «Запит» - аналог об’єкта «View» у базі даних)
- останній – кластерно-спектральна карта показників (дивись таблицю 2)

Таблиця 1.

Об'єкт	*Основне призначення*	*Опис*	*Тип результату*
Зовнішня таблиця	Створити вибірку із заданої БД у вигляді таблиці - кодифікатора	Містить початкові дані типу текстового рядка, в яких описані параметри підключення (паролі доступу, місцезнаходження БД, тіло SQL-запиту	Таблиця
«Основні статистики»	Розрахувати основні статистики заданих полів таблиці – середнє, стандартне відхилення тощо	Містить результат розрахунку	Таблиця
«Запит»	Поєднання за кодами	Аналог аналог обєкта «View» у базі даних	Динамічна таблиця
«Кластерно-спектральна карта»	Групування змінних (та випадків) за мірою «схожості»		

Результат розрахунку представлений на рисунку 3.

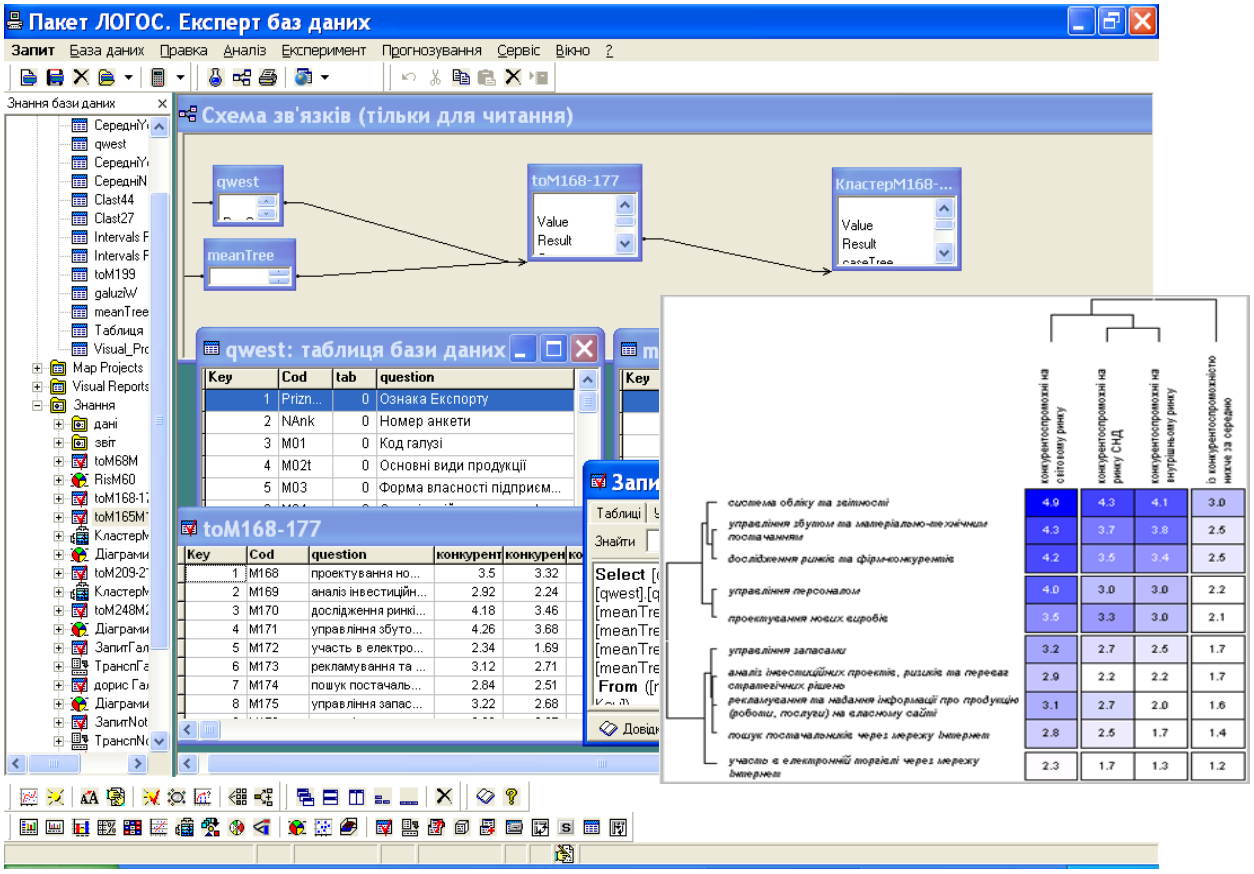

Рис.3. Результат розрахунку ланцюжка методів

Розрахунок виконується автоматично, при чому послідовно виконуються методи, і закінчення виконання попередніх є сигналом для виконання наступних. Таким чином можна відзначити, що така система дозволяє поступове ускладнення методів. Можна одночасно отримувати прогноз поведінки окремих груп (спочатку застосована кластерізація даних, а потім прогноз), можна одночасно отримувати результати розвитку окремих сценаріїв(у вигляді таблиць, після склеювання), обирати найліпший (знаходження оптимуму). Ця система відкрита до поповнення різними об'єктами – методами Data Mining, методами прогностики та оптимізації.

Висновки.

Такий підхід дав можливість створювати комплексні результати послідовного застосування аналітичних методів, а це – прогнозування динамічних рядів з одночасним групуванням за тенденціями, це – візуальний факторний аналіз, в якому групуються компоненти за величиною схожості, це – порівняльний розрахунок різних сценаріїв розвитку подій.

Інженерія аналітичних знань перетворюється в побудову інженером гіпотетичної моделі виходячи з системного підходу до аналізу даних, в якій він вибирає за допомогою графічного інтерфейсу методи обробки даних та знань не поглиблюючись в математичні аспекти методу, отримує результат у вигляді аналітичного знання, поступово ускладнюючи ланцюжок обробки знань.

Прикладом реалізації такого підходу є програмне забезпечення інформаційної та експертно-аналітичної підтримки прийняття рішень Логос-3.0, в якому реалізована аналітична база знань у вигляді файлу з розширенням .sdb. База знань має множину об'єктів, що забезпечують роботу з данними та методами пошуку знань.

Список літератури:

1. Барсегян А.А, Анализ данных и процессов: учеб.пособие/ СПб.:БХВ-Петербург, 2009 – 512с.
2. А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ–Петербург, 2004. – 336 с.
3. Сошникова Л.А.,ич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике. Учеб. Пособие для вузов. – М.:ЮНИТИ-ДАНА, 1999. – 598с.
4. Алгазинов Э.К., Сирота А.А. Системный подход в современных исследованиях. Л. 1978. – 73с.
5. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. Пер. с англ. – М.: Мир, ООО «Издательство АСТ», 2003. – 686с.
6. Джеймс Одел. Агенти і складні системи. – Ж. «Открытые системы», октябрь 2003. – с.54–58

Відомості про автора:

Лисак Володимир Васильович - старший викладач кафедри Інформаційних технологій Київської державної академії водного транспорту (КДАВТ), аспірант, область інтересів - методи аналізу даних (інтелектуального, чисельного, статистичного тощо)

e-mail: vladimir.lysak@gmail.com