УДК 004.891

В.М. ДЖУЛІЙ, О.М. ГОРБАТЮК
Хмельницький національний університет

СТРУКТУРА, ФУНКЦІЇ СИСТЕМИ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ДАНИХ

В статті пропонуються технічні рішення по структурі, функціям та організації інструментарію інтелектуального аналізу даних, заснованого на технології WEB систем. Формулюються стратегічні напрямки розробки програмної системи для підвищення ефективності обробки великих масивів електронних даних для задач інтелектуального аналізу: класифікації, навчання, прогнозування.

Ключові слова: інтелектуальна обробка даних, програмні системи, алгоритми, база даних і знань, інформаційна система.

V.M. DZULIY, A.M. GORBATYUK
Khmelnytsk national university

STRUCTURE, FUNCTION OF INTELLECTUALDATA PROCESSING

This article we offer technical solutions for the structure, function and organization of data mining tools based on WEB technology systems. Formulated strategic directions of development of a software system to improve efficiency of processing large volumes of electronic data mining tasks. classification, training, forecasting.

Keywords: *data mining, software systems, algorithms, database and knowledge information system.*

Вступ. Ефективність роботи сучасного промислового підприємства в інформаційному суспільстві залежить від швидкості і якості задоволення потреб в службовій інформації кожного з працівників. Інформаційні сховища корпоративних інформаційних систем можуть досягати величезних розмірів, що сильно ускладнює пошук. Необхідна інформація часто розподілена по різних інформаційних системам всередині підприємства, її інтеграція утруднена через неоднозначність використовуваної термінології, специфічної структури компонентів інформаційних сховищ, різного рівня компетентності співробітників підприємства.

Великі обсяги інформації привели до вибухового зростання популярності більш широких методів інтелектуального аналізу даних, тому, що інформації стало набагато більше, і вона за самою своєю природою і змістом стає більш різноманітною і обширною. При роботі з великими наборами даних вже недостатньо відносно простої і прямолінійної статистики. Бізнес-вимоги привели від простого пошуку і статистичного аналізу даних до складнішого інтелектуальному аналізу даних. Для вирішення бізнес-завдань потрібно такий аналіз даних, який дозволяє побудувати модель для опису інформації та в кінцевому підсумку призводить до створення результуючого звіту. Цей процес ілюструє рисунок 1.

Процес інтелектуального аналізу даних, пошуку та побудови моделі часто є ітеративним, так як потрібно розшукати і виявити різні відомості, які можна витягнути. Необхідно також розуміти, як зв'язати, перетворити і об'єднати їх з іншими даними для отримання результату. Після виявлення нових елементів і аспектів даних підхід до виявлення джерел і форматів даних з подальшим зіставленням цієї інформації з заданим результатом може змінитися.

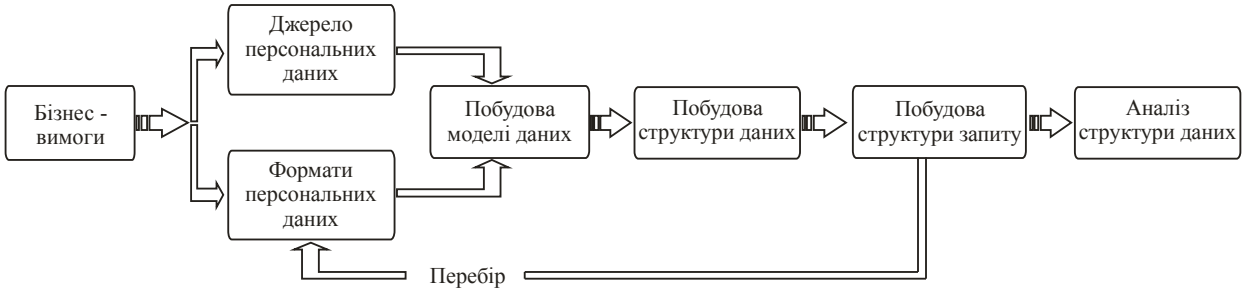

Рис. 1. Процес інтелектуального аналізу даних

Інтелектуальний аналіз даних - це не тільки використовувані інструменти або програмне забезпечення баз даних. Інтелектуальний аналіз даних можна виконати з відносно непотужними системами баз даних і простими інструментами, включаючи створення своїх власних, або з використанням готових пакетів програмного забезпечення. Складний інтелектуальний аналіз даних опирається на досвід і алгоритми, визначені за допомогою існуючого програмного забезпечення та пакетів, причому з різними методами асоціюються різні спеціалізовані інструменти. Більш гнучкий формат зберігання бази документів надає обробці інформації нову спрямованість і ускладнює її. Бази даних SQL строго регламентують структуру і жорстко дотримуються схеми, що спрощує запити до них і аналіз даних з відомими форматом і структурою. Документальні бази даних, які відповідають стандартній структурі типу JSON, або файли з деякою машиночитаючою структурою теж легко обробляти, хоча справа може ускладнюватися різноманітною і мінливою структурою. Наприклад, в Hadoop, який обробляє абсолютно "сирі" дані, може

бути важко виявити і витягти інформацію до початку її обробки та зіставлення.

Методи інтелектуального аналізу даних. Основні методи, які використовуються для інтелектуального аналізу даних, описують тип аналізу і операцію з відновлення даних: асоціація (або відношення), класифікація, кластеризація, прогнозування, послідовні моделі, дерева рішень. Дерево рішень, пов'язане з більшістю інших методів (головним чином, класифікації та прогнозування), можна використовувати або в рамках критеріїв відбору, або для підтримки вибору певних даних в рамках загальної структури. Дерево рішень починають з простого питання, яке має дві відповіді (іноді більше). Кожна відповідь призводить до наступного питання, допомагаючи класифікувати та ідентифікувати дані або робити прогнози. На практиці дуже рідко використовується тільки один з цих методів. Класифікація та кластеризація - подібні методи. Використовуючи кластеризацію для визначення найближчих сусідів, можна додатково уточнити класифікацію. Дерева рішень часто використовуються для побудови і виявлення класифікацій, які можна простежувати на історичних періодах для визначення послідовностей і моделей.

При всіх основних методах часто має сенс записувати і згодом вивчати отриману інформацію. Для деяких методів це абсолютно очевидно. Наприклад, при побудові послідовних моделей і навчанні в цілях прогнозування аналізуються історичні дані з різних джерел та примірників інформації. Інтелектуальний аналіз даних - це не тільки виконання деяких складних запитів до даних, що зберігаються в базі даних. Незалежно від того, чи використовується SQL, бази даних на основі документів, такі як Hadoop, або прості неструктуровані файли, необхідно працювати з даними, форматувати або реструктурувати їх. Потрібно визначити формат інформації, на якому буде грунтуватися метод і аналіз. Потім, коли інформація знаходиться в потрібному форматі, можна застосовувати різні методи (окремо або в сукупності), які не залежать від необхідної базової структури даних або набору даних.

Основна частина. За останні десятиліття відбувся інтенсивний розвиток автоматизованих інформаційних систем, в тому числі в таких областях, як мережеві технології Internet, способи зберігання і представлення знань, мови та інструментарію програмування, методи штучного інтелекту, алгоритми розподілених і хмарних обчислень і т. д.

Науково-технічні досягнення в галузі штучного інтелекту вплинули на формування нових і трансформацію старих класів інформаційних систем - інтелектуальні інформаційні системи, системи інтелектуального аналізу даних, експертні системи, системи підтримки прийняття рішень і ін. На жаль, всі сучасні інструментарії в більшості випадків розрізнені, вирішують частково предметні задачі або підлягають одному класу систем. Разом з тим рівень їх автоматизації дозволяє зробити висновок про можливість розробки «над-системи», інтегруючи в її складі всі найбільш розвинуті інструментарії (підходи, методи, моделі, алгоритми, технології) у вигляді інтелектуального сховища знань і автоматизації процесу інформаційної підтримки прийняття управлінських рішень.

Підвищення ефективності використання накопиченої в електронному вигляді інформації (банків даних, репозитаріїв, баз знань і ін.) шляхом інтеграції та уніфікації форматів збереження і процедур обробки, призвело до розвитку Web- технологій, заснованих на методах інтелектуального аналізу даних (data mining). Таким чином, стає актуальною ідея створення інтегрованої web-системи інтеллектуального сховища знань і автоматизації процесу інформаційної підтримки прийняття управлінських рішень.

Специфіка системи інтелектуальної обробки даних накладає значний відбиток на методологію і технологію її розробки. Технологія створення системи інтелектуальної обробки даних має свої суттєві особливості і відрізняється від процесу проектування і розробки інших інформаційних та програмних систем. Ці відмінності в чималому ступені визначаються тим, що системи інтелектуальної обробки даних представляють собою інтелектуальні інформаційні системи, що базуються на ідеях, принципах і методах штучного інтелекту. Подібні системи вимагають для своєї реалізації спеціалізовані підходи, методи і технології, в значній мірі відмінні від класичних способів розробки програмних систем. Разом з тим постійно зростаючий інтерес до штучного інтелекту сприяє все більш широкому впровадженню і застосуванню інтелектуальних систем в різних прикладних областях. Розробка систем штучного інтелекту (експертно-орієнтованих систем, систем data mining, перекладу, машинного зору і т.д.) поступово виходить на промисловий рівень, набуває рис індустрії. Це призводить до того, що до подібних систем пред'являтися ті ж вимоги, що і до традиційних програмних продуктів.

Специфіка системи інтелектуальної обробки даних полягає в тому, що розробка і реалізація інтелектуальних інформаційних систем досить тривалий і трудомісткий процес, який ще повністю не відпрацьований і нерідко вимагає нових ідей і рішень, вимагає використання неординарних підходів, методів, технологій і засобів. В результаті ефективність і якість одержуваної системи значною мірою визначається талантом і досвідом її розробників. Крім того, проблема реалізації пов'язана ще і з вибором відповідних засобів розробки, які представляють собою окрему складну задачу. Сучасні аналітичні інформаційні системи набувають характерні риси та особливості програмних систем, заснованих на алгоритмах інтелектуального аналізу даних, експертних систем, систем штучного інтелекту та машинного навчання. Виділимо наступні фундаментальні ознаки системи інтелектуальної обробки даних, що визначають її як інструмент експертної підтримки прийняття рішень: здатність до вирішення широкого кола завдань в деякій неформалізованій проблемній області; здатність витягати знання, з даних і представляти їх у вигляді формалізованих моделей знань; моделювання механізмів інтелектуальної діяльності людини; використання знань про предметну область; застосування евристичних методів вирішення задач; здатність

до пояснення отриманого рішення; висока продуктивність, і т.д.

Платформа інтелектуального аналізу даних надає наступні можливості: надається робочий простір (аутентифікація і авторизація, засоби завантаження і редагування файлів); бібліотека готових підсистем аналізу та алгоритмів, в тому числі: класифікація та кластеризація, побудова правил і дерев рішень, нейронні мережі, генетичні алгоритми, статистичні алгоритми і т.д. ; засоби побудови автоматичного колективного рішення на підставі алгоритмів; засоби аналізу ефективності навчання на основі даних; імпорт/експорт вхідних даних і результатів навчання; візуалізація результатів аналізу.

До стратегічних планів відноситься можливість розширення функцій користувача по створенню та коригуванню індивідуальних розділів бібліотеки алгоритмів, в тому числі: редактор для складання нових алгоритмів на основі метамови; можливість викладати алгоритми в публічний доступ; можливість будувати сценарії з використанням алгоритмів аналізу; інтеграція з бізнес-процесами користувачів/замовників (сервер автоматично по завершенні аналізу або навчання буде відправляти спеціально сконфігуровані дані на сервер замовника, тим самим, буде організований автоматичний вивід даних в систему замовника). робота з сервером (автоматичне введення даних в систему інтелектуальної обробки даних).

Система інтелектуальної обрботкі даних представляє собою програмно-апаратний комплекс, що налаштовується на використання різних класів розпізнавання, мов і форматів представлення даних і знань, апарату синтезу та аналізу модельних представлень, а також орієнтована на надання послуг у сфері вирішення наукових і прикладних задач обробки різноманітних даних. Комплекс інтегрує в своєму складі високопродуктивне апаратне забезпечення (у тому числі потужні обчислювальні сервери, обширні дискові сховища) і ефективне програмне забезпечення, призначене для вирішення широкого спектру завдань інтелектуального аналізу даних та допускає налаштування і розширення підтримуваних функцій.

Система інтелектуальної обробки даних характеризується наступними особливостями: доступ до системи забезпечується за допомогою мережі Інтернет. Більшість функцій системи може використовуватися через веб-інтерфейс, що працює з браузером. При необхідності роботи з конфіденційною інформацією в системі забезпечується шифрування переданих і збережених даних (зокрема, для доступу може бути використаний протокол HTTPS); основна частина послуг і ресурсів системи надається тільки для зареєстрованих користувачів. Деякі з ресурсів системи, наприклад загальновідомі інформаційно-довідкові матеріали, відкриті для широкого доступу і не вимагають реєстрації для використання; зареєстровані користувачі отримують можливість доступу до всього спектру послуг і ресурсів системи. Набір послуг, що надаються і ресурсів, а також їх якісні та кількісні характеристики та обмеження визначаються відповідно до статусу користувача в системі і можуть з часом змінюватися залежно від його вимог і завдань; робота в рамках системи здійснюється відповідно до концепції «проектів». Для рішень своїх задач користувач за допомогою програмного забезпечення системи організовує (створює) один або кілька проектів. Для кожного проекту користувач може орендувати і використовувати (в рамках існуючих обмежень) послуги та ресурси, необхідні йому для розв'язання відповідних задач. Автор (власник) проекту може надати до проекту доступ іншим користувачам або їх групам і при необхідності надати кожному користувачу або групі набір прав, що визначають можливості і допустимі операції при роботі над цим проектом. Дії всіх користувачів, що працюють в рамках проекта, протоколюються в системі і можуть бути надалі проаналізовані і в ряді випадків скасовані; система інтелектуальної обробки даних надає програмний інтерфейс, що дозволяє стороннім додаткам і веб-сайтам взаємодіяти з системою по протоколу HTTP і використовувати її ресурси за певними процедурами; користувачі, яким потрібно підвищений рівень безпеки та/або які з яких-небудь причин не хочуть або не можуть передавати і зберігати в системі свої дані, мають можливість отримати і встановити на своїх комп'ютерах спеціальне клієнтське програмне забезпечення, що дозволяє в автономному режимі (без підключення до мережі Інтернет) використовувати основні функції системи інтелектуальної обробки даних і виконувати обробку та аналіз даних локально. При наявності з'єднання з Інтернет клієнтський додаток на вимогу і під керуванням користувача може здійснювати повну або часткову синхронізацію своїх проектів з системою; система є розширюваною в аспекті підтримуваних функцій і, відповідно, адаптується під різні класи розв'язуваних завдань. До складу системи входять інструментальні і програмні інтерфейси (API), що дозволяють її користувачам додавати нові функції і модулі та робити їх доступними для інших користувачів.

Основні функції системи інтелектуальної обробки даних: система надає засоби для вирішення задач з наступних областей: статистична обробка даних (кореляційний і регресійний аналіз, дисперсійний аналіз і т. п.), розпізнавання образів (класифікація з навчанням), кластеризація (класифікація без навчання), ідентифікація (виявлення розпізнавальних ознак досліджуваних об'єктів), прогнозування (визначення тенденцій розвитку процесів), витяг знань з даних (data mining) і текстів (text mining). Завдяки відкритій архітектурі і API системи набір підтримуваних класів задач може бути достатньо просто доповнений, в тому числі за рахунок використання модулів сторонніх розробників; система підтримує імпорт даних з різних джерел, в тому числі з текстових файлів у форматі CSV, XML і HTML, з електронних таблиць в форматі Excel і OpenDocument, з реляційних баз даних, а також з веб-додатків і служб, таких як таблиці Google, бази даних Zoho Creator або служби Яндекса. Крім того, дані можуть вводитися, змінюватися та редагуватися безпосередньо через інтерфейс системи. Експорт даних і результатів може бути виконаний в файли різних форматів, зокрема CSV, XML, HTML, PDF, RTF, Excel, JPEG, PNG. Крім зазначених, в систему можуть бути інтегровані інші конвертери імпорту/ експорту, що забезпечують роботу зі специфічними форматами і

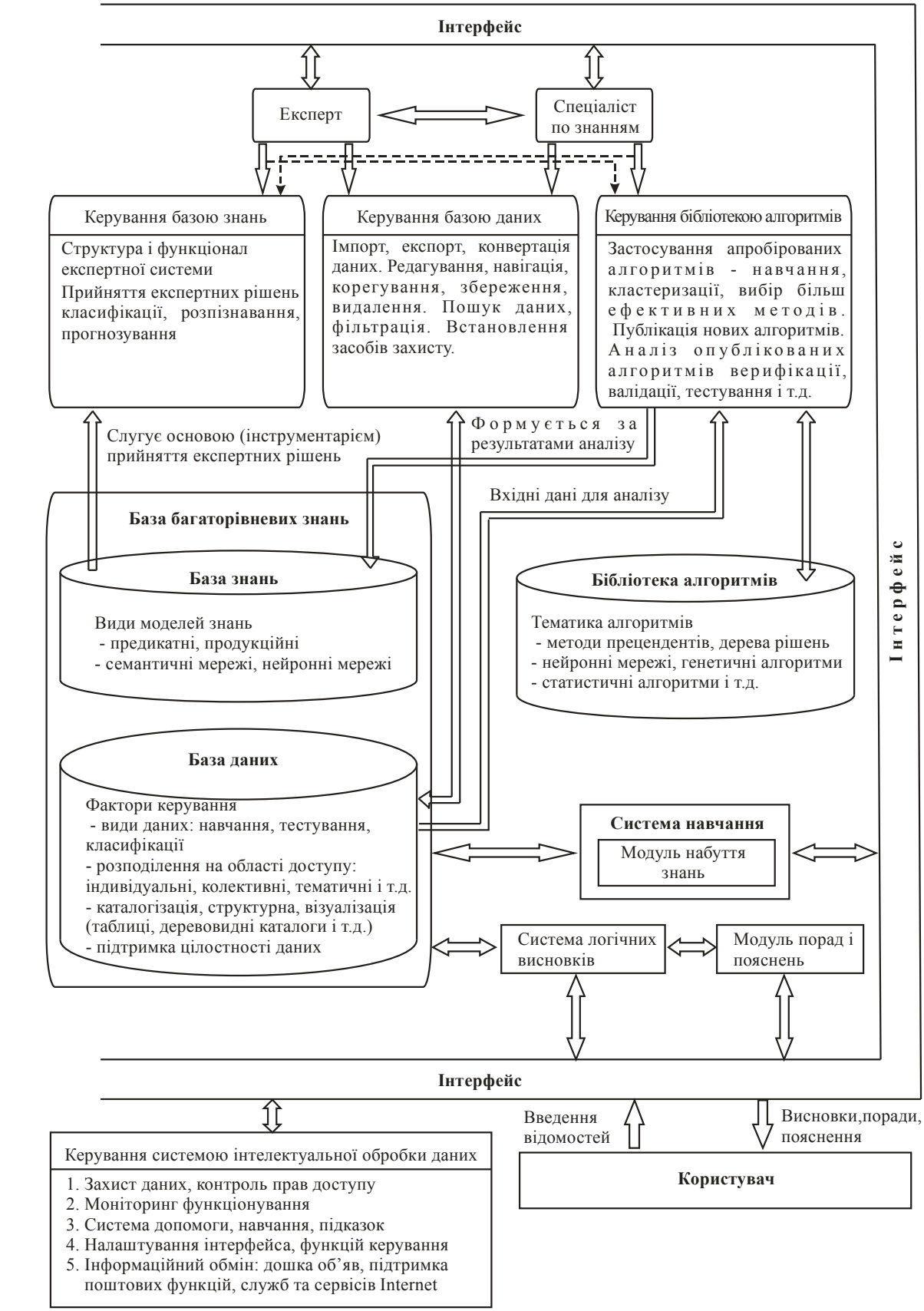

Рис. 2. Структурна схема системи інтелектуальної обробки даних

джерелами даних; система містить інструменти, що забезпечують можливості для візуалізації та графічного представлення вихідних даних і результатів їх обробки в різних формах, у тому числі у вигляді графіків і діаграм, а також для формування різноманітних звітів, які можуть бути опубліковані в рамках системи або експортовані для подальшого використання поза системою; система включає інтерактивні підручники, довідники та модулі тестування з тематики інтелектуальної обробки даних, покликані навчити користувачів ефективному рішенню відповідних задач за допомогою системи, а також підвищити їх рівень знань щодо

моделей і методів обробки даних. Крім того, система надає можливості для організації на її базі електронних навчальних курсів та проведення автоматичного тестування знань для різних предметних областей; для кожного користувача система створює веб-сайт, функціонуючий в рамках системи і доступний іншим користувачам системи. На своєму сайті користувач може агрегатувати і публікувати будь-які матеріали, що не суперечать прийнятим в системі правилами, і при необхідності обмежувати до них доступ. Крім змісту, користувачі в певних межах можуть змінювати оформлення і структуру своїх сайтів; система підтримує різні комунікаційні сервіси для спілкування користувачів між собою і отримання зворотного зв'язку від них, до числа яких належать обмін електронними повідомленнями (у тому числі чат в режимі реального часу) і списки розсилки, підписка на новини та інформування про зміни що відбуваються в системі, тематичні форуми та дискусії, проведення опитувань і голосувань.

Система управління інтелектуальною обробкою даних - це програмний комплекс, що дозволяє автоматизувати процес управління як сайтом в цілому, так і елементами в рамках сайту: макетами сторінок, шаблонами виведення даних, структурою, інформаційним наповненням, користувачами і правами доступу, а також додатковими сервісами: списки розсилки, ведення статистики, пошук, засоби взаємодії з користувачами і т. д.

Проектні рішення в рамках розробки інтелектуального сховища знань і автоматизації процесу інформаційної підтримки прийняття управлінських рішень разом з технологіями інтелектуального аналізу даних, методами штучного інтелекту, моделями подання інформації, електронними базами по різних предметних областях слугують інструментом для підвищення ефективності наукової, інноваційної та освітньої діяльності, а також в промисловості, в економіці, в медицині і т. д.

Практична реалізація проекту дозволить використовувати його як самоналаштовуючу, адаптивну, відкриту інтелектуальну інформаційну систему з інтегрованими функціями експертної системи і підсистеми інтелектуального аналізу даних.

Структурна схема системи інтелектуальної обробки даних представлена на рисунку 2.

Запропонований математичний апарат і методика його використання в процесі аналізу даних є інструментом наукової оцінки та формального обґрунтування прийнятих інженерних рішень в різних областях прогнозування, класифікації, прийняття рішень.

Висновки. Незалежно від того, чи використовується SQL, бази даних на основі документів, такі як Hadoop, або прості неструктуровані файли, необхідно працювати з даними, форматувати або реструктурувати їх. Потрібно визначити формат інформації, на якому буде ґрунтуватися метод і аналіз. Потім, коли інформація знаходиться в потрібному форматі, можна застосовувати різні методи (окремо або в сукупності), які не залежать від необхідної базової структури даних або набору даних.

Пропонуються технічні рішення по структурі, функціям та організації інструментарію інтелектуального аналізу даних, заснованого на технології WEB систем. Формулюються стратегічні напрямки розробки програмної системи для підвищення ефективності обробки великих масивів електронних даних для задач інтелектуального аналізу, класифікації, навчання, прогнозування.

Література

1. Гаскаров Д. В. Интеллектуальные информационные системы; Интеллектуальная информационная технология, Экспертные системы: Учеб. пособие / Д. В. Гаскаров, Д. В. Сикулер, В, В. Фомин, И. К. Фомина. СПб,; СПГУВК, 2004. – 362 с.

2. Попов Э. В. Искусственный интеллект: В 3 ки. Ки. I. Системы общения и экспертные системы: Справочник / Под ред. Э. В. Попова. М: Радио и связь, 1990.- 464 с.

3. Бриллюэн Л. Наука и теория информации / Бриллюэн Л.; [пер. с фр. Е.В. Гайдукова и Н.Н. Родман] – М.: Физматгиз, 1960. – 749 с.

4. Марманис Х. Алгоритмы интеллектуального Интернета / Х. Марманис, Д.Бабенко [пер. с англ], СПб.; Символ- Плюс, 2011.- 480с.

5. Фомин В. В. Автоматизация логического моделирования программного обеспечения с применением формального аппарата семиотических систем. / В. В.Фомин СПб.: Энергоатомиздат, Санкт-Петербургское отделение, 2000.- 250 с.

References

1. Gaskarov DV Intelligent Information Systems; Intelligent information technology, Expert Systems: Proc. Manual / DV Gaskarov, DV Sikuler, B, Fomin, IK Fomin. St. Petersburg,; SPGUVK, 2004 - 362 p.

2. Popov EV Artificial Intelligence: In 3 ki. Key. I. communication systems and expert systems: Directory / Ed. E. Popov. M: Radio and communication, 1990.- 464 p.

3. L. Brillouin, Science and Information Theory / Brillouin L .; [trans. with fr. EV Gaidukova and NN Rodman] - M .: Fizmatgiz, 1960 - 749 p.

4. Marmanis X. Mining Algorithms Internet / X. Marmanis, D.Babenko [trans. from English], St. Petersburg .; Symbol- Plus, 2011.- 480p.

5. Fomin VV Automation logic simulation software prime¬neniem formal apparatus of semiotic systems. / V. Fomin Petersburg .: Energoatomizdat, St. Petersburg Department, 2000.- 250 p.

Рецензія/Peer review : 3.9.2014 р. Надрукована/Printed : 16.11.2014 р.