УДК 004.652

О. М. Верес, Р. М. Оливко
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

КЛАСИФІКАЦІЯ МЕТОДІВ АНАЛІЗУ ВЕЛИКИХ ДАНИХ

Описано особливості класифікації методів і технологій аналітики Великих даних, групи методів і технологій аналітики Великих даних, які класифікуються з урахуванням функціональних зв’язків та формальної моделі цієї інформаційної технології. Розв’язано задачу визначення концептів онтології аналітики Великих даних.

Ключові слова: аналіз, Великі дані, візуалізація даних, модель, Data Mining, Text Mining, MapReduce.

This article describes the features of classification methods and technologies, analytics Big data. Described group of methods and technologies, analytics Big data that are graded according to the functional relationships and formal model of information technology. The problem of the definition of ontology concepts analytics Big data.

Key words: analysis, Big data, visualization data, model, Data Mining, Text Mining, MapReduce.

Вступ. Загальна постановка проблеми

Великі дані дають змогу побачити і зрозуміти зв’язки між фрагментами інформації, які донедавна ми тільки намагалися вловити [1]. У зв’язку зі швидким поширенням розумних і взаємопов’язаних пристроїв і систем обсяг зібраних даних зростає загрозливими темпами. У деяких галузях близько 90 % даних зберігаються в неструктурованому вигляді, а їх обсяг збільшується на 50 % щорічно. Що стосується аналізу великих даних та інших аналітичних завдань, поточні рішення не забезпечують швидкість реакції системи, необхідну для роботи із завданнями аналізу, що знижує продуктивність користувача і затягує процес прийняття рішень [2].

Змінюються методи ведення бізнесу. Змінюється поведінка споживачів. Змінюються самі споживачі. Для збереження конкурентоспроможності підприємства прагнуть в реальному часі дізнаватися, коли клієнти щось купують, де вони купують, і навіть що вони думають перед тим, як зайти в магазин або відвідати Web-сайт. Допомогу в цьому можуть надати Великі дані, аналіз Великих даних та інтегрована платформа для бізнес-аналітики (BI) і аналізу Великих даних [1–4].

Аналіз останніх досліджень і публікацій

Стандартна бізнес-практика великомасштабного аналізу даних ґрунтується на понятті “корпоративного сховища даних” (*Enterprise Data Warehouse, EDW*), запити до якого надходять від програмного забезпечення “бізнес-аналітики” (*Business Intelligence, BI*) [5]. Інструменти BI дають змогу створювати звіти та інтерактивні інтерфейси, узагальнення даних за допомогою агрегатних функцій (наприклад, обчислити кількість або середнє) до різноманітних розподілів ієрархічних даних на групи.

Традиційно вважається, що ретельно спроектоване сховище даних відіграє центральну роль у разі правильного застосування інформаційних технологій. Сховище даних традиційно контролюють спеціально призначені працівники IT, які не тільки супроводжують систему, а й ретельно контролюють доступ до неї, щоб керівні особи могли гарантовано розраховувати на високий рівень обслуговування [5].

Кількість внутрішньокорпоративних великомасштабних джерел даних істотно зростає: великі бази даних сьогодні виникають навіть на основі єдиного джерела потоків даних про відвідування Web-сайтів (*click-stream*), журналів програмних систем, архівів електронної пошти і форумів тощо. Загальновизнаною стала значущість аналізу даних. Численні компанії демонструють, що складний аналіз даних сприяє зменшенню витрат та навіть прямому зростанню доходів. Результатом цих можливостей є масовий перехід до збирання та використання даних у декількох організаційних одиницях корпорацій.

У цьому змінному кліматі збирання розрізнених великомасштабних даних доцільним є підхід, який називають *могутнім аналізом даних* (МАД; ***Magnetic, Agile, Deep (MAD) data analysis***) [5]. Акронім MAD походить від трьох аспектів цього середовища, що відрізняють його від ортодоксальних сховищ даних, а саме: Магнетична (***magnetic***); Гнучкість (***agile***); Ґрунтовність (***deep***).

Великі дані (*англ*. ***Big data***) – серія підходів, інструментів і методів опрацювання структурованих та неструктурованих даних величезних обсягів і значного різноманіття для отримання зрозумілих для людини результатів, ефективних в умовах безперервного приросту, розподілу по численних вузлах обчислювальної мережі, що сформувалися в кінці 2000-х років, альтернативних традиційним системам управління базами даних і рішень класу Business Intelligence [7]. Є три типи завдань, пов'язаних з Великими даними (Big Data) [1–4, 6, 7].

1) *зберігання і управління*. Обсяг даних в сотні терабайт або петабайт не дає змоги легко зберігати їх та керувати ними за допомогою традиційних реляційних баз даних;

2) *неструктурована інформація*. Більшість Великих даних неструктуровані;

3) *аналіз Великих даних*. Як аналізувати неструктуровану інформацію? Як на основі Великих даних складати прості звіти, будувати та впроваджувати поглиблені прогностичні моделі?

Робота з Великими даними не схожа на звичайний процес бізнес-аналітики, коли просте додавання відомих значень приносить результат. Працюючи з великими даними, результат одержують, очищаючи їх за допомогою послідовного моделювання: спочатку висувається гіпотеза, будується статистична, візуальна або семантична модель, на її підставі перевіряється достовірність висунутої гіпотези і потім пропонується наступна. Цей процес вимагає від дослідника або інтерпретації візуальних значень, або складання інтерактивних запитів на основі знань, або розроблення адаптивних алгоритмів "машинного навчання", здатних отримати потрібний результат. Причому час життя такого алгоритму може бути доволі коротким [2, 6].

Не вирішені раніше частини загальної проблеми

Розроблення проекту корпоративної СППР з керування даними передбачає виникнення певних складнощів, що пов'язані з Великими даними. Треба знайти нові підходи до аналізу даних і, за необхідності, наявні методи повинні бути розширені. Це місце, де математичні науки можуть дати значний внесок: будівля на фундаменті поточних статистичних методів і виявлення нових методів, щоб збільшити або замінити старі, які менш придатні, роблячи аналітику ефективною, а найголовніше, переконавшись, що правильні висновки отримують з наявних даних. До етапу розроблення чи використання засобів аналітики Великих даних треба дослідити технології та підходи для подолання складнощів отримання значущих знань зі структурованих та неструктурованих даних, з акцентом на застосування інформаційної технології Великі дані.

Цілі (завдання) статті

За інтенсивного розвитку бізнесу для збереження конкурентоспроможності підприємства та опрацювання значних обсягів накопичених структурованих та неструктурованих даних допомогти може інформаційна технологія Великі дані. Актуальним є застосування методів і технологій аналізу Великих даних та інтегрованої платформи для бізнес-аналітики. Метою роботи є дослідження особливостей класифікації методів і технологій аналітики Великих даних з урахуванням означення та особливостей застосування технології Великих даних.

Описання методів і технологій аналітики Великих даних (Big Data Analytics)

Формальна модель великих даних як інформаційної технології така [8–15]:

$$\mathbf{BD} = \langle \mathbf{Vol}_{BD}, \mathbf{Ip}, \mathbf{A}_{BD}, \mathbf{T}_{BD} \rangle,$$

де $\mathbf{Vol}_{BD}$ – множина типів обсягів; $\mathbf{Ip}$ – множина типів джерел даних (інформаційних продуктів); $\mathbf{A}_{BD}$ – множина методик аналізу Великих даних; $\mathbf{T}_{BD}$ – множина технологій обробки Великих даних.

На основі означення Великих даних [9] можна сформулювати основні принципи роботи з такими даними: горизонтальна масштабованість; стійкість до відмов; локальність даних. Усі сучасні засоби роботи з Великими даними так чи інакше відповідають цим трьом принципам. Для того, щоб їх дотримуватися, необхідно придумувати якісь методи, способи і парадигми розроблення засобів опрацювання даних.

Сьогодні наявна множина $\mathbf{A}_{BD} = \{\mathbf{A}_i\}$ різноманітних методик аналізу масивів даних, в основу яких покладено інструментарій, запозичений з статистики та інформатики.

Необхідність у нових засобах для аналізу обґрунтована тим, що даних стає більше, більше їх зовнішніх і внутрішніх джерел, тепер вони складніші та різноманітніші (структуровані, неструктуровані та слабкоструктуровані), використовуються різні схеми індексації (реляційні, багатовимірні, noSQL). Колишні способи опрацювання даних вже неефективні – *Big Data Analytics* поширюється на великі й складні масиви, тому ще використовують терміни *Discovery Analytics* (аналітика, що відкриває) і *Exploratory Analytics* (аналітика, що пояснює).

Сьогодні не розмежовують вживання термінів Big Data і Big Data Analytics. Ці терміни описують як самі дані, так і технології управління та методи аналізу [16, с. 13].

Big Data Analytics є розвитком концепції Data Mining. Ті самі завдання, сфери застосування, джерела даних, методи і технології. За роки, що минули з моменту появи концепції Data Mining до настання ери Великих даних, революційно змінилися обсяги даних, що аналізуються, з'явилися системи високопродуктивних обчислень, нові технології, зокрема MapReduce і її численні програмні реалізації. З появою соціальних мереж постали і нові завдання.

Data Mining – це процес підтримки ухвалення рішень, що ґрунтується на пошуку в сирих даних прихованих закономірностей, раніше невідомих, нетривіальних, практично корисних та доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності [16–18]. Data Mining – це особливий підхід до аналізу даних. Акцент робиться не тільки на добуванні фактів, а й на генерації гіпотез.

Якщо підхід DataMining доповнити технологією MapReduce і вимогою 4V (Volume (обсяг), Velocity (швидкість), Variety (різноманітність), Veracity (достовірність), то це відобразить функціональні зв'язки Big Data Analytics (рис. 1).

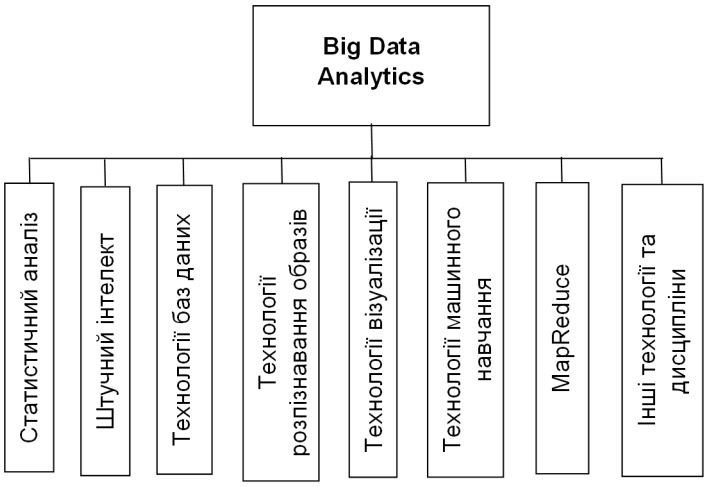

Рис. 1. Функціональні зв'язки аналітики Великих даних

Аналіз великих обсягів даних і необхідності зрозуміти значення з індивідуальної поведінки потребує методів оброблення, які виходять за межі традиційних статистичних методів [16].

Методики і методи аналізу, які застосовують до великих даних, також описано в звіті McKinsey [19, с. 27–31]: методи DataMining; краудсорсинг; консолідація та інтеграція даних; машинне навчання; нейронні мережі, мережевий аналіз, оптимізація, зокрема, генетичні алгоритми; розпізнавання образів; аналітика, прогнозування; імітаційне моделювання; просторовий аналіз; статистичний аналіз; візуалізація аналітичних даних.

BothManyika(2011) [19] і Chen (2012) запропонували такий список методів аналітики Великих даних (в алфавітній послідовності): A/B тестування (A/Btesting), правило навчання асоціації (Association rule learning), класифікація (Classification), кластерний аналіз (Cluster analysis), злиття і інтеграція даних (Data fusion and data integration), Ансамблі навчання (Ensemble learning), генетичні алгоритми (Genetic algorithms), машинного навчання (Machine learning), обробки природної мови (Natural Language Processing), Нейронні мережі (Neural networks), мережевий аналіз (Network analysis), розпізнавання образів (Pattern recognition), Прогнозне моделювання (Predictive modelling),

регресія (Regression), Настроїв аналіз (Sentiment Analysis), Обробка сигналів (Signal Processing), Просторовий аналіз (Spatial analysis), статистика (Statistics), кероване і некероване навчання (Supervised and Unsupervised learning), моделювання (Simulation), аналіз часових рядів та візуалізації (Timeseries analysis and Visualization).

Опишемо групи методів і технологій аналітики Великих даних, які класифікуються з урахуванням функціональних зв'язків та формальної моделі цієї інформаційної технології, а саме: методи Data Mining, технології Text Mining, технологія MapReduce, візуалізація даних, інші технології та методики аналізу (рис. 2).

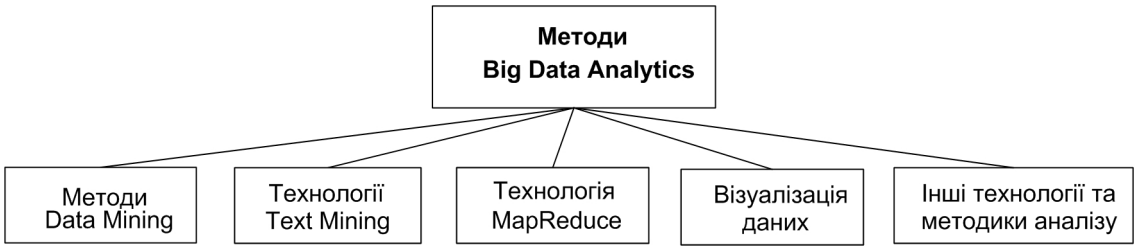

Рис. 2. Групи методів аналітики Великих даних

Методи інтелектуального аналізу даних (Data Mining). Застосування методів і технологій Data Mining дає змогу розв'язати такі задачі [16–18, 20–24]: класифікація (*Classification*); кластеризація (*Clustering*); асоціація (*Associations*); послідовність (*Sequence*), або послідовна асоціація (*sequential association*); прогнозування (*Forecasting*); визначення відхилень (*Deviation Detection*), аналіз відхилень або викидів; оцінювання (*Estimation*); аналіз зв'язків (*Link Analysis*); візуалізація (*Visualization, Graph Mining*); підбивання підсумків (*Summarization*) – опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Методи Data Mining поділяють на дві групи: навчання з учителем (*Supervised Learning*); навчання без учителя (*Unsupervised Learning*) [16–18]. Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні й кібернетичні методи. Ця схема поділу ґрунтується на різних підходах щодо навчання математичним моделям [16–18, 20–24].

Опишемо найпридатніші з них для аналізу Великих даних [16–18, 20–33].

Асоціативні правила (*Association Rule Learning*). Набір методик для виявлення взаємозв'язків, тобто асоціативних правил, між змінними величинами у великих масивах даних. Для аналізу ринкового кошика застосовують **аналіз прихованих закономірностей** (A*ssociation Analysis*).

Класифікація (*Classification*). Набір методик, які дають змогу передбачити поведінку споживачів у певному сегменті ринку (прийняття рішень про покупку, відтік, обсяг споживання тощо).

Метод дерев рішень (*Decision Trees*) є одним з найпопулярніших методів розв'язання завдань класифікації та прогнозування. У найпростішому вигляді дерево рішень – це спосіб подання правил в ієрархічній, послідовній структурі. Метод дерев рішень зазвичай називають "наївним" підходом.

Кластерний аналіз (*Cluster Analysis*). Статистичний метод класифікації об'єктів за групами у результаті виявлення наперед не відомих загальних ознак. Приклад – сегментування ринку.

Для вирішення завдання кластеризації на графах застосовують алгоритм Girvanand Newman методу MLP (Markov Cluster Algorithm).

Для аналізу Великих багатовимірних даних розроблено методологію "Dynamic Quantum Clustering" (DQC), що реалізує парадигму пошуку як "нехай дані говорять про себе самі" [32]. Метод DQC (як і багато інших методів аналітики Великих даних) "працює" без попереднього знання про ті "структури", їх тип і топології, які можуть бути "приховані" в даних і виявлені в результаті його застосування. Метод добре працює з багатовимірними даними і час аналізу лінійно залежить від розмірності.

Регресія (*Regression*). Набір статистичних методів для виявлення закономірності між зміною залежної змінної та однієї або декількох незалежних.

Аналіз часових рядів (*Time Series Aanalysis*). Набір запозичених зі статистики та цифрової обробки сигналів методів аналізу повторюваних з плином часу послідовностей даних. **Аналіз викидів** (*Outlieran Aalysis*) застосовують для виявлення шахрайства, особистого маркетингу, медичного аналізу.

Машинне навчання (*Machine Learning*). Напрям в інформатиці (історично за ним закріпилася назва "штучний інтеллект"), який має на меті створення алгоритмів самонавчання на основі аналізу емпіричних даних. Машинне навчання сьогодні використовується: для розпізнавання спаму або не спаму повідомлень електронної пошти; для отримання знань про переваги користувача та надання рекомендацій, що ґрунтуються на цій інформації; для визначення кращого контенту для залучення потенційних клієнтів; для встановлення ймовірності виграшу справи та відповідності юридичним нормам пред'явлених рахунків.

Кероване і некероване навчання (*Supervised and Unsupervised Learning*). Набір методик, що ґрунтуються на технологіях машинного навчання, які дають змогу виявити функціональні взаємозв'язки в аналізованих масивах даних. Некероване навчання має спільні риси з кластерним аналізом.

Ансамблі навчання (*Ensemble Learning*). У цьому методі задіється множина предикативних моделей, за рахунок чого поліпшується якість прогнозів.

Еволюційні алгоритми, генетичні алгоритми (*Evolution Analysis, Genetic Algorithms*). Генетичні алгоритми навіяні природою еволюційних процесів – тобто таких механізмів, як успадкування, мутації та природний добір. Ці механізми використовуються для "еволюціонування" корисного вирішення проблем, які потребують оптимізації. У цій методиці можливі рішення подають у вигляді "хромосом", які можуть комбінуватися і мутувати. Як і в процесі природної еволюції, виживає найпристосованіша особина.

Нейронні мережі (*Neural Networks*) – це клас моделей, що ґрунтуються на аналогії з роботою мозку людини та призначені для розв'язання різноманітних задач аналізу даних після проходження етапу навчання на даних. За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажів, показники фінансового ринку, розпізнавати сигнали, розробляти самонавчальні системи.

Візуалізація даних

Візуалізація (*Visualization*). Методи графічного подання результатів аналізу великих даних у вигляді діаграм або анімації для спрощення інтерпретації, полегшення розуміння отриманих результатів. Візуалізація аналітичних даних – зображення інформації у вигляді рисунків, графіків, схем і діаграм з використанням інтерактивних можливостей та анімації для результатів, а також вихідних даних для подальшого аналізу [33, с. 173–210].

Наочне представлення результатів аналізу Великих даних має принципове значення для їхньої інтерпретації [34–41]. Сприйняття людини обмежене, і вчені продовжують вести дослідження у галузі вдосконалення сучасних методів подання даних у вигляді зображень, діаграм або анімацій. Новими прогресивними методами візуалізації є: хмара тегів; кластерограма; історичний потік; просторовий потік.

Технології Text Mining. Підґрунтям технології **Text Mining** – статистичний та лінгвістичний аналіз, методи штучного інтелекту. Ця технологія застосовується для проведення аналізу, забезпечення навігації та пошуку в неструктурованих текстах [42–47]. Застосування інформаційних систем класу Text Mining дає змогу користувачам набувати нових знань.

Технології Text Mining – набір методів, які призначені для видобування відомостей з текстів на основі сучасних ІКТ, що дає змогу виявити закономірності, які забезпечують користувачам отримання корисних даних та нових знань. Основна мета Text Mining – надати аналітику можливість працювати з великими обсягами початкових даних за рахунок автоматизації процесу здобуття потрібних даних.

Основними методами технології Text Mining є: класифікація (*classification*); кластеризація (*clustering*); побудова семантичних мереж або аналіз зв'язків (*Relationship, Event andFact Extraction*); здобуття феноменів, фактів, понять (*feature extraction*); автоматичне реферування, створення анотацій (*summarization*); відповідь на запити (*question answering*); тематичне індексування (*thematic indexing*); пошук за ключовими словами (*keyword searching*); засоби підтримки та створення таксономії (*oftaxonomies*) і тезаурусів (*thesauri*).

Прикладом ефективного застосовування технологій Text Mining є проведення контент-аналізу. **Контент-аналіз** (*Content Analysis*) – це якісно-кількісне, систематичне опрацювання, оцінювання та інтерпретація форми і змісту тексту.

Інші технології та методики досліджень

Опишемо декілька технологій і дисциплін дослідження даних з погляду технології Великих даних [20–33].

A/B тестування (*A/B testing, Splittesting*). Методика маркетингового дослідження, в якій контрольна вибірка по черзі порівнюється з іншими. Метод використовується для оптимізації Web-сторінок відповідно до заданої мети.

Обробка природної мови (*Natural Language Processing (NLP*). Набір запозичених з інформатики та лінгвістики методик розпізнавання природної мови людини.

Аналіз настроїв (*Sentiment Analysis*). В основу методик оцінки настроїв споживачів покладено технології розпізнавання природної мови людини. Аналіз настроїв допомагає дослідникам визначити настрої спікерів або авторів щодо теми.

Мережевий аналіз (*Network Analysis*). Набір методик аналізу зв'язків між вузлами в мережах. Стосовно соціальних мереж дає змогу аналізувати взаємозв'язок між окремими користувачами, компаніями, спільнотами тощо.

Оптимізація (*Optimization*). Набір числових методів для редизайну складних систем і процесів для поліпшення одного або декількох показників. Допомагає у прийнятті стратегічних рішень, наприклад, складу виведеної на ринок продуктової лінійки, у проведенні інвестиційного аналізу тощо.

Розпізнавання образів (*Pattern Recognition*). Набір методик з елементами самонавчання для передбачення поведінкової моделі споживачів.

Прогнозне моделювання (*Predictive Modeling*). Набір методик, які дають змогу створити математичну модель наперед заданого ймовірного сценарію розвитку подій.

Обробка сигналів (*Signal Processing*). Запозичений з радіотехніки набір методик, який має на меті розпізнавання сигналу на тлі шуму і його подальшого аналізу.

Просторовий аналіз (*Spatial Analysis*). **Просторовий аналіз** – використання топологічної, геометричної та географічної інформації в даних. Набір частково запозичених зі статистики методик аналізу даних. Джерелом великих даних у цьому випадку є геоінформаційні системи (ГІС).

Статистика (*Statistics*). Наука про збирання, організацію та інтерпретацію даних, зокрема розроблення опитувальників і проведення експериментів. Статистичні методи часто застосовують для оцінкових суджень про взаємозв'язки між тими чи іншими подіями.

Моделювання (*Simulation*). Моделювання поведінки складних систем часто використовується для прогнозування, передбачення і опрацювання різних сценаріїв під час планування.

Краудсорсинг (*Crowdsourcing*). Методика збирання даних з великої кількості джерел. Краудсорсинг – категоризація та збагачення даних силами широкого, невизначеного кола осіб, з метою використання їхніх творчих здібностей, знань і досвіду із застосуванням інформаційно-комунікаційних технологій.

Злиття та інтеграція даних (*Data Fusion and Data Integration*). Набір технік, що дають змогу інтегрувати різнорідні дані з різноманітних джерел інформації для проведення глибинного аналізу. Цей набір методик дає змогу аналізувати коментарі користувачів соціальних мереж і зіставляти з результатами продажів у режимі реального часу.

Технологія MapReduce. Створення і підтримка сховищ даних обсягом в терабайт, петабайт і більше уможливилась завдяки технологіям розподілених файлових систем [48]. Розподілені системи опрацювання даних, замість зберігання даних в одній файловій системі, зберігають та індексують дані на декількох (навіть тисячах) жорстких дисках і серверах. Створюється також "карта" (*map*), на якій міститься інформація про місцезнаходження тих чи інших даних. Однією з найвідоміших систем, що використовують цей підхід, є **Hadoop**. Щоб опрацювати дані в розподіленій файловій системі, необхідно виконувати низькорівневі обчислення, такі як підсумовування, агрегування тощо, в місці їхнього фізичного розміщення в розподіленій файловій системі. Створити карту (*map*) виконаних обчислювальних алгоритмів і відстежувати локальні

результати, а потім акумулювати результати (reduced). Цей підхід і шаблон проведення обчислювальних алгоритмів отримав назву **MapReduce** [48–53]. MapReduce – це фреймворк для обчислення деяких наборів розподілених завдань з використанням великої кількості комп'ютерів ("нод"), що утворюють кластер. Опрацьовуватися можуть дані, які зберігаються або в файловій системі (неструктуровано), або в базі даних (структуровано).

Багато практичних завдань можна реалізувати у цій моделі програмування. Є безліч інструментів для проведення такого агрегування даних у розподіленій файловій системі, що дає змогу легко здійснювати цей аналітичний процес.

Наведений опис методів і технологій аналізу Великих даних дає змогу побудувати онтологію відповідно до підходу METHONTOLOGY [54-57], який відображає процес ітеративного проектування. За методологією METHONTOLOGY глосарій термінів містить всі терміни (концепти та їхні екземпляри, атрибути, дії), важливі для аналізу Великих даних, і їхні природно-мовні описи.

Глосарій термінів онтології аналізу Великих даних містить означені вище терміни, які можна семантично розділити на три групи: структура завдання (групи технологій аналітики, зв'язки), дані, що наповнюють задачу (методи, що застосовують для кожної групи), і результати обчислень (рекомендації щодо використання Великих даних для підвищення ефективності ухвалення рішень). Онтологія аналізу Великих даних розроблена засобами Protégé-OWL.

Висновки

Великі дані мають вагоме практичне значення як технологія, призначена для вирішення актуальних повсякденних проблем, але породжує ще більше нових. Великі дані здатні змінити наш спосіб життя, праці й мислення.

Однією з умов успішного розвитку світової економіки на сучасному етапі стає можливість фіксувати й аналізувати величезні масиви і потоки інформації. Є думка, що країни, які оволодіють найефективнішими методами роботи з Великими даними, чекає нова індустріальна революція. Напрям "Big Data" концентрує зусилля в організації зберігання, оброблення, аналізу величезних масивів даних.

У результаті проведених досліджень, з використанням розробленої формальної моделі інформаційної технології Великі дані, обґрунтовано поділ на групи методів і технологій аналітики Великих даних. Для досягнення поставленої мети запропоновано, з урахуванням функціональних зв'язків та формальної моделі цієї інформаційної технології Великі дані, класифікувати всі методики так: методи Data Mining, технології Text Mining, технологія MapReduce, візуалізація даних, інші технології та методики аналізу. Подано описання характеристик та особливостей методів і технологій, що належать до кожної з виділених груп, враховуючи означення Великих даних.

Отже, використовуючи розроблену формальну модель та результати критичного аналізу методів і технологій аналізу Великих даних, можна побудувати онтологію аналізу Великих даних.

Подальші роботи стосуватимуться дослідження методів, моделей та інструментів для удосконалення онтології аналітики Великих даних та ефективнішої підтримки розроблення структурних елементів моделі системи підтримки прийняття рішень з керування Великими даними.

1. Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. – М. : Манн, Иванов и Фербер, 2014. – 240 с. 2. Большие данные и аналитика [Електронний ресурс]. – Режим доступу: http://www-03.ibm.com/systems/ru/technicalcomputing/bigdata.html 3. Агеева А. Аналитики предупредили об опасности больших данных [Електронний ресурс] / Анна Агеева. – Режим доступу: http://bigdata.cnews.ru/news/top/2015-10-23_eksperty_predosteregayut_ot_nepravilnogo_obrashcheniya. 4. Названы причины торможения рынка больших данных [Електронний ресурс]. – Режим доступу: http://bigdata.cnews.ru/news/top/2015-11-20_analitiki _otsenili_tempy_rosta_mirovogo_rynka. 5. Коэн Дж. МОГучие способности: новые приемы анализа больших данных [Електронний ресурс] / Джеффри Коэн, Брайен Долэн, Марк Данлэп, Джозеф Хеллерстейн, Кейлэб Велтон; пер. с англ. Сергей Кузнецов. – Режим доступу: http://citforum.ru/database/articles/mad_skills/

6. History and evolution of big data analytics [Електронний ресурс]. – Режим доступу: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html 7. Mitchell R. 8 big trends in big data analytics [Електронний ресурс] / Robert L. Mitchell // Computerworld, OCT 23, 2014. – Режим доступу : http://www.computerworld.com/article/ 2690856/big-data/8-big-trends-in-big-data-analytics.html 8. Большие данные (Big Data) [Електронний ресурс]. – Режим доступу: http://tadviser.ru/a/125096. 9. Inmon W. H. Big Data – getting it right: A checklist to evaluate your environment / [Електронний ресурс] / W. H. Inmon. // DSSResources.COM, – 2014. – Режим доступу: http://dssresources.com/papers/features/ inmon/inmon01162014.htm. 10. Шаховська Н. Б. Організація великих даних у розподіленому середовищі / Н. Б. Шаховська, Ю. Я. Болюбаш, О. М. Верес // Обчислювальна техніка та автоматизація: [зб. наук. пр. ДонНТУ]. – Донецьк, 2014. – С. 147–155. – (Вісник / ДонНТУ ; № 2 (27). 11. Shakhovska N. B. Big Data Federated Repository Model / N. B. Shakhovska, Yu. Ja. Bolubash, O. M. Veres // The Experience of Designing and Application of CAD Systems in Microelectronics (CADMS'2015) Proc. ofthe XIII-thInt. Conf., (Polyana-Svalyava (Zakarpattya), Ukraine, 24-27 February, 2015). – Lviv: Publishing Lviv Polytechnic, 2015. – P. 382–384. 12. Veres O. Elements of the Formal Model Big Data / Oleh Veres, Natalya Shakhovska // Перспективні технології і методи проектування МЕМС: матеріали XI міжнар. конф. MEMSTECH'2015, 2–6 вересня 2015, Львів / Нац. ун-т "Львів. політехніка". – Львів: Вид-во Львів. політехніки, 2015. – С. 81–83. 13. Shakhovska N. Data space architecture for Big Data managering / N. Shakhovska, O. Veres, Y. Bolubash, L. Bychkovska-Lipinska // Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT'2015). – P. 184–187, Lviv, 2015. DOI: 10.1109/STC-CSIT.2015.7325461 14. Shakhovska N. Generalized formal model of Big Data / N. Shakhovska, O. Veres and M. Hirnyak, // ECONTECHMOD: an international quarterly journal on economics of technology and modelling processes, vol. 5, no. 2, 2016. – P. 33–38. 15. Shakhovska N. Big Data Information Technology and Data Space Architecture / N. Shakhovska, O. Veres, Y. Bolubash // Sensors & Transducers, vol. 195, no. 12. P. 69–76, 2015. 16. Барсегян А. А. Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд. перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с. 17. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям (+ CD) / Н. Б. Паклин, В. И. Орешков. – СПб.: Питер, 2009. – 624 с. 18. Дюк В. Data Mining: учебный курс (+CD) / В. Дюк, А. Самойленко. – СПб.: Питер, 2001. – 368 с. 19. Manyika J. Big data: The next frontier for innovation, competition, and productivity / Manyika James. Mc Kinsey Global Institute, June, 2011. – 156 с. 20. Журавлёв Ю. И. Распознавание. Математические методы. Программная система. Практические применения / Ю. И. Журавлёв, В. В. Рязанов, О. В. Сенько. – М. : Фазис, 2006. – 176 с. 21. Зиновьев А. Ю. Визуализация многомерных данных / А. Ю. Зиновьев. – Красноярск: Изд. Красноярского гос. техн. ун-та, 2000. – 180 с. 22. Чубукова И. А. Data Mining: учеб. пособ. / И. А. Чубукова. – М. : Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с. 23. Ситник В. Ф. Інтелектуальний аналіз даних (дейтамайнінг): навч. посіб. / В. Ф. Ситник, М. Т. Краснюк. – К.: КНЕУ, 2007. – 376 с. 24. Ian H. Witten. Data Mining: Practical Machine Learning Tools and Techniques / Ian H. Witten, Eibe Frank, Mark A. Hall. – 3rd Edition. – Morgan Kaufmann, 2011. – 664 с. 25. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance / Bernard Marr. – John Wiley&Sons Ltd, 2015. – 256 с. 26. Einav L. The Data Revolution and Economic Analysis [Електронний ресурс] / Liran Einav, Jonathan Levin // NBER Working PaperNo. 19035, 2013. – Режим доступу : http://www.nber.org/chapters/ c12942.pdf 27. Ваняшин А. За большими данными следит ПАНДА / А. Ваняшин, А. Климентов, В. Кореньков // Суперкомпьютеры. 2013. – № 3 (11). – С. 56–61 28. Серов Д. Аналитика "больших данных" – новые перспективы [Електронний ресурс] / Денис Серов // "StorageNews", №1 (49), 2012. – Режим доступу : http://www.storagenews.ru/49/EMC_BigData_49.pdf 29. Ronen Sh. Links that speak: The global language network and its association with global fame [Електронний ресурс] / Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, César A. Hidalgo // PNAS, Vol. 111, No. 52, 2014. – Режим доступу : http://stevenpinker.com/files/pinker/files /pnas_hildago_et_al_global_language_ network_2014.pdf 30. Aflalo Y. Spectral multidimensional scaling [Електронний ресурс] / Yonathan Aflalo, Ron Kimmel // PNAS, vol. 110, no. 45, November 5, 2013. – Режим доступу : http://www.cs.technion.ac.il/~ron/PAPERS/Journal/AflaloKimmelPNAS2013.pdf 31. Gadepally V. Big Data Dimensional Analysis [Електронний ресурс] / Vijay Gadepally, Jeremy Kepner. arXiv:1408.0517v1. –

Режим доступу : https://arxiv.org/pdf/1408.0517v1.pdf 32. Weinstein M. Analyzing Big Data with Dynamic Quantum Clustering *[Електронний ресурс] / M. Weinstein, F. Meirer, A. Hume, Ph. Sciau, G. Shaked, R. Hofstetter, E. Persi, A. Mehta, D. Horn. arXiv:1310.2700.– Режим доступу :* https://arxiv.org/ftp/arxiv/papers/1310/1310.2700.pdf. *33. Паклин, Н. Б. Бизнес-аналитика: от данных к знаниям [Текст] : учеб. пособ. / Н. Б. Паклин, В. И. Орешков. – 2-е изд., испр. — СПб. : Питер, 2013. – 702 с. 34. Желязны Д. Говори на языке диаграмм : пособие по визуальным коммуникациям для руководителей / Д. Желязны. – М. : Институт комплексных стратегических исследований, 2004. – 220 с. 35. Роэм Д. Практика визуального мышления. Оригинальный метод решения сложных проблем / Д. Роэм. – М . : Манн, Иванов и Фербер, 2014. – 396 с. 36. Тафти Э. Представление информации [Електронний ресурс] / Э. Тафти. – Режим доступу :* http://envisioninginformation.daiquiri.ru/15 *37. Яу Н. Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами / Н. Яу. – М. : Манн, Иванов и Фербер, 2013. – 352 с. 38. Iliinsky N. Designing Data Visualizations / N. Iliinsky, J. Steele. – Sebastopol :O'Reilly, 2011. – 110 с. 39. Krum R. Cool infographics: effective communication with datavisualization and design / R. Krum. – Indianapolis: Wiley, 2014. – 348 с. 40. Тьюки Дж. Анализ результатов наблюдений: разведочный анализ / Дж. Тьюки; под ред. В. Ф. Писаренко. – М.: Мир, 1981. – 693 с. 41. Alper C. New Software for Visualizing the Past, Presentand Future [Електронний ресурс] / C. Alper, K. Brown, G. R. Wagner // DSSResources.COM, 09/23/2006. – Режим доступу :*http://dssresources.com/papers/ features/alperbrown&wagner/alperbrown&wagner 9212006.html *42. Барсегян А. А. Анализ данных и процессов: учеб. пособ. / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – – 3-е изд. перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с. 43. Text Mining [Електронний ресурс]. – Режим доступу:* http:/ /statsoft.ru/home/textbook/modules/sttextmin. html#index *44. Ландэ Д. Глубинный анализ текстов: технология эффективного анализа текстовых данных [Електронний ресурс] / Дмитрий Ландэ. – Режим доступу:* http://visti.net/~dwl/art/dz/ *45. Барсегян А. А. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд. перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с. 46. Линючев П. Text Mining: современные технологии на информационных рудниках [Електронний ресурс] / Павел Линючев // PC Week/RE №6 (564), 27 февраля – 5 марта 2007. – Режим доступу до ресурсу:*https://www.pcweek. ru/idea/article/detail.php?ID=82081 *47. Плескач В. Л. Інформаційні системи і технології на підприємствах : підручник / В. Л. Плескач, Т. Г. Затонацька. – К. : Знання, 2011. – 718 с. 48. Стоунбрейкер М. MapReduce и параллельные СУБД: друзья или враги? [Електронний ресурс] / Майкл Стоунбрейкер, Дэниэль Абади, Дэвит Девитт, Сэм Мэдден, Эрик Паулсон, Эндрю Павло, Александр Разин ; пер. с англ. Сергей Кузнецов // Communications of the ACM, vol. 53, no. 1, January 2010. – Режим доступу:* http://citforum.ru/database/articles/mr_vs_dbms-2/ *49. Березин А. Map-Reduce на примере MongoDB [Електронний ресурс] / Антон Березин. – Режим доступа:* https://habrahabr.ru/post/184130/ *50. Лебеденко Е. Технология GoogleMapReduce: разделяй и властвуй [Електронний ресурс] / Евгений Лебеденко. – Режим доступа :*http://www.computerra.ru/ 82659/mapreduce/ *51. Павло Э. Сравнение подходов к крупномасштабному анализу данных [Електронний ресурс] / Эндрю Павло, Эрик Паулсон, Александр Разин, Дэниэль Абади, Дэвид Девитт, Сэмюэль Мэдден, Майкл Стоунбрейкер; пер. с англ. Сергей Кузнецов. – Режим доступа :* http://citforum.ru/database/articles/mr_vs_dbms/2.shtml *52. BigData от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Електронний ресурс] . – Режим доступа:* https://habrahabr.ru/company/dca/blog/267361/ *53. Big data от А до Я. Часть 3: Приемы и стратегии разработки MapReduce-приложений [Електронний ресурс]. – Режим доступа:* https://habrahabr.ru/company/dca/blog/270453/ *54. Гаврилова Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб. : Питер, 2000. – 384 с. 55. Гаврилова Т. А. Онтология для изучения инженерии знаний // Труды Международной научно-практической конференции KDS-2001. – 2001. 56. Гаврилова Т. А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем // Новости искусственного интеллекта. – 2003. – № 2. – С. 24–30. 57. Литвин В. В. Бази знань інтелектуальних систем підтримки прийняття рішень: монографія / В. В. Литвин; Міністерство освіти і науки, молоді та спорту України, Національний університет "Львівська політехніка". – Львів : Вид-во Львівської політехніки, 2011. – 240 с.*