DOI: 10.15587/2312-8372.2018.123612

ОНТОЛОГІЯ АНАЛІЗУ BIG DATA

Литвин В. В., Висоцька В. А., Верес О. М., Бродяк О. Я., Орищин О. Г.

1. Вступ

При інтенсивному розвитку бізнесу, для збереження конкурентоспроможності підприємства та опрацювання значних обсягів накопичених структурованих та неструктурованих даних, допомогу надає інформаційна технологія (IT) Big Data (BD). Актуальним є застосування методів і технологій аналізу BD та інтегрованої платформи для Business Intelligence. BD дають змогу побачити і зрозуміти зв'язки між фрагментами інформації. Це безліч нових завдань громадської безпеки, глобальних економічних моделей, недоторканності приватного життя, усталених моральних правил, правових відносин людини, бізнесу та держави.

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 понад 50 % щорічно. Для збереження конкурентоспроможності, інновацій і швидкого виведення продуктів і послуг на ринок необхідно мати можливість аналізувати ці дані і отримувати на їх основі аналітичну інформацію швидко та економічно. Що стосується аналізу BD і інших аналітичних завдань, поточні рішення не забезпечують швидкість реакції інформаційної системи (IC), необхідну для роботи із завданнями аналізу, що знижує продуктивність користувача і затягує процес прийняття рішень [1, 2]. Змінюються споживачі, змінюється світ бізнесу. Сьогодні вже недостатньо дослідження тільки даних про продажі. Мета розгортання інтегрованої платформи для Business Intelligence (BI) і аналізу BD полягає в тому, щоб копати глибше і краще розуміти – чому, де, що і як – про клієнтів, продукти і компанії. 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 надають BD, аналіз BD та інтегрована платформа для BI і аналізу BD [1–4].

2. Обєкт дослідження та його технологічний аудит

Обєктом данного дослідження є процеси аналізу Big Data.

На вході системи є методи та IT аналізу BD, які описні в [1–5]. На виході системи є онтологічна модель правил аналізу BD $O=<X,R,F>$. Таксономія понять онтології X задає методику аналізу Big Data A_{BD} (кореневе поняття онтології). Оптимальне визначення множини відношень між цими поняттями R та множини правил F аналізу BD, формалізованих за допомогою дескриптивної логіки DL, дозволить ефективно опрацьовувати BD, тобто: S: RABD$\rightarrow O$.

Одним з найбільш проблемних місць є відсутність чіткої класифікації методів аналізу BD, наявність якої значно полегшить вибір оптимального та ефективного алгоритму аналізу цих даних в залежності від їх структури. Аналіз BD

мають вагоме практичне значення для сучасних ІТ та слугує вирішенню актуальних повсякденних проблем, але при цьому породжує ще більше нових. Ефективний та своєчасний аналі BD здатний змінити наш спосіб життя, праці і мислення. Однією з умов успішного розвитку світової економіки на сучасному етапі стає можливість фіксувати і аналізувати величезні масиви і потоки інформації. Країни, які оволодіють найбільш ефективними методами роботи з BD, чекає нова індустріальна революція. Напрямок «Big Data» концентрує зусилля в організації зберігання, опрацювання, аналізу величезних масивів даних. Поширеною помилкою навколо великих обсягів даних є очікування, що придбання потужної комп'ютерної інфраструктури негайно забезпечить переваги для бізнесу, замість того, щоб ІТ, інформатика і математична наука йшли рука в руку. Інфраструктура є необхідною, але отримання користі від великих обсягів даних вимагає також застосування більш складних методів їх аналізу.

3. Мета та задачі дослідження

Мета дослідження – розроблення програмної системи *S* формалізації правил аналізу Big Data RABD у вигляді онтологічної бази знань (БЗ) з метою її використання для опрацювання та аналізу будь-яких BD.

Для досягнення поставленої мети необхідно:

1. Дослідити особливості класифікації методів і технологій аналітики Big Data з врахуванням означення та особливості застосування відповідних ІТ.
2. Розробити формальну модель аналізу BD.
3. Розробити онтологічну БЗ аналізу BD.
4. Побудувати правила аналізу Big Data RABD.

4. Дослідження існуючих рішень проблеми

Стандартна бізнес-практика великомасштабного аналізу даних ґрунтується на понятті EDW (Enterprise Data Warehouse), запити до якого надходять від програмного забезпечення (ПЗ) BI [1–5]. Інструменти BI дають змогу створювати звіти та інтерактивні інтерфейси, узагальнювати дані за допомогою агрегатних функцій до різноманітних розподілів ієрархічних даних на групи.

Ретельно спроектоване EDW відіграє центральну роль при правильному застосуванні ІТ. Проектування та еволюція детальної схеми сховища знань (СД) є загальним принципом дисциплінованої інтеграції даних великих підприємств, удосконалюючи результати та подання всіх бізнес-процесів. Результуюча база даних (БД) відіграє роль репозиторію характеристик критичних бізнес-функцій. Крім того, сервер БД, що зберігає СД, традиційно є основним обчислювальним засобом, який слугує центральним, масштабуючим механізмом ключової корпоративної аналітики. Концептуальне та обчислювальне центральне місце СД робить його критично важливим дорогим ресурсом, який використовують для продукування звітів з великою кількістю даних. І ці звіти орієнтують на керівних осіб, які ухвалюють рішення. СД традиційно контролюється спеціально призначеними співробітниками ІТ, які не тільки супроводжують ІС, а й ретельно контролюють доступ до неї, щоб керівні особи могли гарантовано розраховувати на високий рівень обслуговування [5].

Хоча в багатьох ситуаціях цей ортодоксальний підхід СД продовжує застосовуватися, ряд факторів сприяє просуванню зовсім іншої філософії управління великомасштабними даними на підприємствах. Зберігання даних тепер обходиться настільки дешево, що невеликі підгрупи підприємства можуть розробити окрему БД астрономічного масштабу в межах свого власного бюджету. Кількість внутрішньо-корпоративних великомасштабних джерел даних значно зростає: великі БД сьогодні виникають навіть на основі єдиного джерела потоків даних про відвідування Web-сайтів (*click-stream*), журналів ІС, архівів е-пошти і форумів тощо. Загальновизнаною стає значимість аналізу даних. Численні компанії демонструють, що складний аналіз даних сприяє скороченню витрат та навіть прямого зростання доходів. Результатом цих можливостей є масовий перехід до збирання та використання даних в декількох організаційних одиницях корпорацій. Перевага цього переходу полягає у підвищенні ефективності та зростанні культури використання даних, але він посилює децентралізацію даних, з якої покликане боротися СД. У цьому змінному кліматі збирання розрізнених великомасштабних даних доцільним є підхід MAD (Magnetic, Agile, Deep data analysis) [5].

У сучасному аналізі BD використовують все більш складні статистичні методи, що далеко виходять за межі узагальнення (*rollup*) і деталізації (*drilldown*) традиційних методів BI. При виконанні цих алгоритмів аналітикам часто потрібно досліджувати величезні набори даних, не вдаючись до використання зразків і вибірок. Сучасне СД має служити і ґрунтовним (глибоким) репозиторієм даних, і механізмом підтримки виконання складних алгоритмів. Сьогодні є зростаюча потреба в могутніх аналітиках даних. Часто вони є висококваліфікованими статистиками, що володіють хорошими знаннями в області ПЗ, але зазвичай фокусуються на ґрунтовному аналізі даних, а не на управлінні БД. Для підтримки їхньої діяльності потрібно застосовувати підхід MAD до проектування СД та створення інфраструктури систем БД. При досягненні даних цілей виникають важливі проблеми вибору методів та ІТ для аналізу BD. Робота з BD не подібна на звичайний процес BI, де просте додавання відомих значень приносить результат. При роботі з BD результат виходить в процесі їхнього очищення шляхом послідовного моделювання: спочатку висувається гіпотеза, будується статистична, візуальна або семантична модель, на її підставі перевіряється достовірність висунутої гіпотези і потім висувається наступна. Цей процес вимагає від дослідника або інтерпретації візуальних значень, або складання інтерактивних запитів на основі знань, або розроблення адаптивних алгоритмів ML, здатних отримати потрібний результат. Причому час життя такого алгоритму часто досить короткий [2, 6]. Є п’ять основних підходів до аналізу BD [7]:

1. *Discovery інструменти* корисні впродовж життєвого циклу інформації для швидкого, інтуїтивного вивчення та аналізу інформації, отриманої з будь-якої комбінації структурованих і неструктурованих джерел. Дані додатки дають можливість аналізу джерел даних поряд з традиційними системами BI. Відсутнє попереднє моделювання, користувачі швидко залучають нові ідеї, формують значущі висновки, і приймають обґрунтовані рішення.

2. *Інструменти BI* мають важливе значення для звітності, аналізу та управління ефективністю, в першу чергу з транзакційних даних зі СД та ІС виробництва. Додатки забезпечують широкі можливості для BI та управління ефективністю.

3. *In-DatabaseAnalytics* – методи для пошуку шаблонів і відношень в даних. Застосовують в БД, відсутнє переміщення даних з інших аналітичних серверів, що прискорює цикл опрацювання інформації та зменшує сукупну вартість.

4. *Hadoop* – попереднє опрацювання даних для трендів макро ідентичності або знаходження елементів даних значення OUTOF-діапазону. Організації використовують Hadoop як прекурсор для форм аналітики.

5. *Управління рішеннями* – прогнозне моделювання, бізнес-правила та самонавчання для прийняття обґрунтованого рішення на основі поточного контексту. Створює процеси прийняття рішень в режимі реального часу.

Всі ці підходи застосовуються для виявлення прихованих взаємозв'язків.

Сьогодні немає відмінності у вживанні термінів Big Data і Big Data Analytics. Ці терміни описують як самі дані, так і технології управління та методи аналізу [8–10]. Big Data Analytics є розвитком концепції Data Mining. Одні і ті ж завдання, сфери застосування, джерела даних, методи та ІТ. З моменту появи концепції Data Mining до настання ери BD революційним чином змінилися обсяги даних, що аналізуються, з'явилися ІС високопродуктивних обчислень, нові ІТ, в тому числі MapReduce та її численне ПЗ. З появою соціальних мереж з'явилися і нові завдання. Data Mining є процесом 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 [10–12].

Data Mining є особливим підходом до аналізу даних. Акцент робиться не тільки на добуванні фактів, а й на генерації гіпотез. Створені в процесі гіпотези необхідно перевіряти за допомогою звичайного аналізу в рамках звичних схем і/або зі залученням експертів предметної області (ПО). В даному підході використовують традиційні інструменти аналізу, такі як математична статистика (регресійний, кореляційний, кластерний, факторний аналіз, аналіз часових рядів, дерева рішень тощо). А також ті інструменти, що пов'язані зі штучним інтелектом (ШІ) (ML, нейронні мережі, генетичні алгоритми, нечітка логіка тощо). 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. Аналіз великих обсягів даних і необхідності зрозуміти значення з індивідуальної поведінки вимагають методів опрацювання, які виходять за рамки традиційних статистичних методів [10–13]. В [13] запропоновний список методик і методів аналізу BD, який не претендує на повноту, проте в ньому відображені найбільш затребувані в різних галузях підходи. Крім того, деякі з BD даних і можуть з успіхом використовуватися для менших за обсягом масивів (наприклад, A/B-тестування, регресійний аналіз). Безумовно, чим більший і диверсифікований масив піддається аналізу, тим точніші та релевантні дані вдається отримати на виході.

5. Методи досліджень

Big Data – серія підходів, інструментів і методів опрацювання структурованих та неструктурованих даних величезних обсягів. Це також джерело значного різноманіття для отримання зрозумілих людиною результатів, ефективних в умовах безперервного приросту, розподілу по вузлах мережі, альтернативних традиційним системам управління БД і рішень класу BI [7]. Є три типи завдань пов'язаних з BD [1–4, 6, 7]: зберігання і управління, опрацювання неструктурованої інформації, аналіз BD (рис. 1).

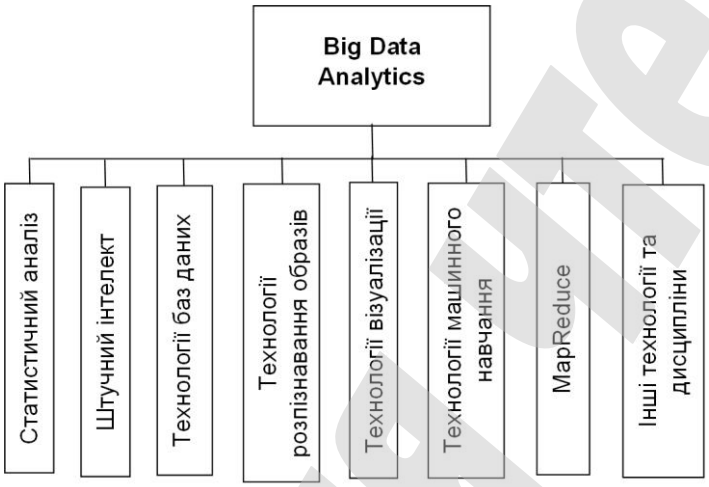

Рис. 1. Функціональні зв'язки аналітики Big Data

Формальна модель BD як IT має таке подання [8, 9]:

$$BD=<Vol_{BD}, Ip, A_{BD}, T_{BD}>, \quad (1)$$

де Vol_{BD} – множина типів обсягів; Ip – множина типів джерел даних (інформаційних продуктів); A_{BD} – множина методик аналізу Big Data; T_{BD} – множина технологій опрацювання Big Data.

Виходячи з означення BD [9], можна сформулювати основні принципи роботи з такими даними: горизонтальна масштабованість, стійкість до відмов та локальність даних. Усі сучасні засоби роботи з BD 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 $A_{BD}=\{A_i\}$ різноманітних методик аналізу масивів даних, в основі яких лежить інструментарій, запозичений з статистики та інформатики (рис. 2, 3).

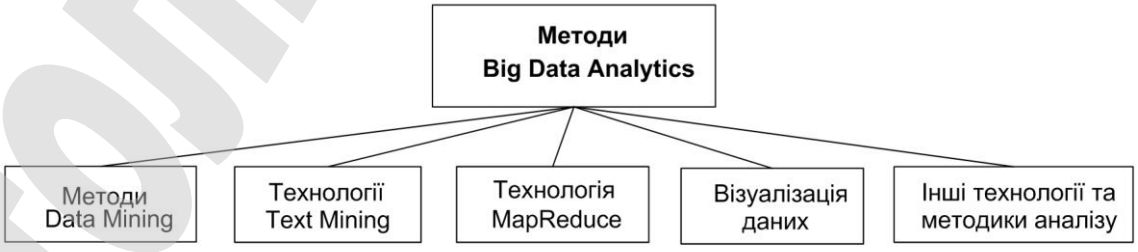

Рис. 2. Групи методів аналітики Big Data

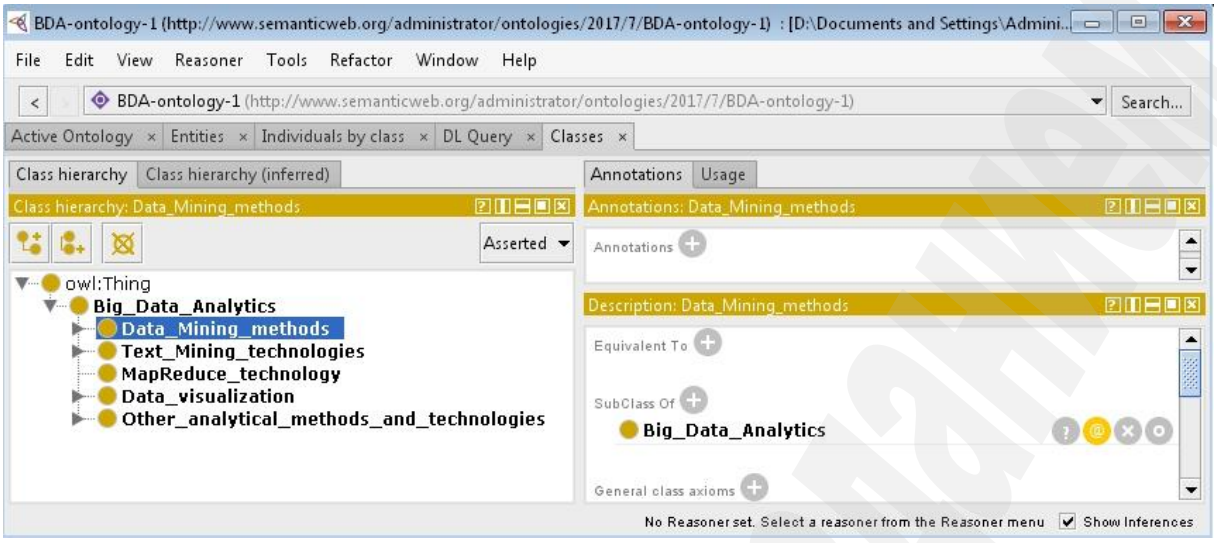

Рис. 3. Підкласи класу Big Data analysis засобами Protégé 3.4.7

Групи методів та технологій для аналізу BD формально подамо у вигляді кортежу:

$$A_{BD}=\langle M_{Data\ Mining}, M_{Machine\ Learning}, M_{Visualization}, T_{Text\ Mining}, T_{MapReduce}, T_{other}, K_{BD}, fdm, fml, fmv, fmt, fmr, fmo\rangle, \quad (2)$$

де $M_{Data\ Mining}$ – набір методики видобування даних (Data Mining);

$M_{Machine\ Learning}$ – набір методики Machine Learning;

$M_{Visualization}$ – методи графічного подання аналізу BD;

$T_{Text\ Mining}$ – технології Text Mining;

$T_{MapReduce}$ – технології MapReduce;

T_{other} – інші конкретні методи та технології для аналізу в BD;

fdm – функція визначення методики Data Mining відповідно до типу задачі;

fml – функція визначення методики Machine Learning відповідно до типу задачі;

fmv – функція визначення методики графічного подання аналізу BD відповідно до типу задачі;

fmt – функція визначення технології Text Mining відповідно до типу задачі;

fmr – функція визначення технології MapReduce відповідно до типу задачі;

fmo – функція визначення іншої методики аналізу BD відповідно до типу задачі.

Причому

$$K_U \subset K_{BD} \text{ та } K_{BD}=K_{Data\ Mining}\bigcup K_{Machine\ Learning}\bigcup K_{Visualization}\bigcup \bigcup K_{Text\ Mining}\bigcup K_{MapReduce}\bigcup K_{other}, \quad (3)$$

де K_{BD} – критерії та параметри аналізу BD;

K_U – критерії та параметри аналізу конкретної BD;

$K_{Data\ Mining}$ – критерії та параметри вибору методики Data Mining відповідно K_U;

$K_{Machine\ Learning}$ – критерії та параметри вибору методики Machine Learning відповідно K_U;

$K_{Visualization}$ – критерії та параметри вибору методики Visualization відповідно K_U;

$K_{Text\ Mining}$ – критерії та параметри вибору технології Text Mining відповідно K_U;

$K_{MapReduce}$ – критерії та параметри вибору технології MapReduce відповідно K_U;

K_{other} – критерії та параметри вибору іншої методики аналізу BD відповідно K_U.

Необхідність в нових засобах для аналізу BD обґрунтована тим, що даних стає більше, більше їх зовнішніх і внутрішніх джерел, тепер вони складніші та різноманітніші (структуровані, неструктуровані та слабо структуровані), використовуються різні схеми індексації (реляційні, багатовимірні, noSQL). Колишні способи опрацювання даних є неефективними – *Big Data Analytics* поширюється на великі і складні масиви, в тому числі *Discovery Analytics* і *Exploratory Analytics*.

Data Mining або інтелектуальний аналіз даних є виявленням прихованих взаємозв'язків або закономірностей між змінними у великих масивах неопрацьованих даних. Вибір методу Data Mining для аналізу BD залежить від типу задачі. Відповідно, згідно (2) $M_{\text{Data Mining}}$ подамо як кортеж:

$$M_{Data\ Mining}=\langle Tk_{Data\ Mining}, Md_{Data\ Mining}, fdm, K_{Data\ Mining}\rangle, \quad (4)$$

де $Tk_{Data\ Mining}$ – задачі видобування даних (Data Mining) при $Tk_{Data\ Mining}=fdm(Md_{Data\ Mining}, K_{Data\ Mining})$; $Md_{Data\ Mining}$ – методи видобування даних (Data Mining).

Застосування методів Data Mining дає змогу розв'язати такі задачі [14–18]:

$$Tk_{Data\ Mining}=\langle T_{Classification}, T_{Clustering}, T_{Associations}, T_{Sequence}, T_{Forecasting}, T_{Deviation\ Detection}, T_{Estimation}, T_{LinkAnalysis}, T_{Graph\ Mining}, T_{Summarization}\rangle, \quad (5)$$

де $T_{Classification}$ – виявлення ознак, які описують групи об'єктів наборів досліджуваних даних – класи; за даними ознаками новий об'єкт належатиме до того чи іншого класу;

$T_{Clustering}$ – кластеризація (поділ) об'єктів на групи;

$T_{Associations}$ – знаходження закономірностей між пов'язаними подіями у наборі даних;

$T_{Sequence}$ – виявлення взаємозв'язку між пов'язаними у часі подіями (послідовність вирізняється високою ймовірністю ланцюжка пов'язаних у часі подій);

$T_{Forecasting}$ – на ґрунті особливих властивостей накопичених даних оцінюються майбутні значення показників;

$T_{Deviation\ Detection}$ – виявлення й аналіз даних, що найбільше відрізняються від загальної чисельності даних, виявлення нехарактерних шаблонів;

$T_{Estimation}$ – прогноз безперервних значень ознак;

$T_{LinkAnalysis}$ – знаходження залежностей у наборі даних;

$T_{\text{Graph Mining}}$ – створення графічного образу аналізованих даних для ілюстрації наявності закономірностей в даних;

$T_{Summarization}$ – опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Так, відповідно до (3), (4) для розв'язання $T_{Classification}=ftc(MT_{Data\ Mining})$ використовують:

$$MT_{Data\ Mining}=\langle M_{Nearest\ Neighbor}, M_{k\text{-}Nearest\ Neighbor}, M_{Bayesian\ Networks}, M_{Tree}, M_{Neural\ Networks}\rangle, \quad (6)$$

де $M_{Nearest\ Neighbor}$ – метод найближчих сусідів для класифікації даних;

$M_{k\ Nearest\ Neighbor}$ – метод k-Nearest Neighbor для класифікації даних;

$M_{Bayesian\ Networks}$ – Bayesian Networks для класифікації даних;

$M_{Neural\ Networks}$ – Neural Network для класифікації даних;

M_{Tree} – індукція дерев рішень для класифікації даних;

ftc – функція визначення методу Data Mining для задачі класифікації.

Найвідоміший алгоритм розв'язку $T_{Associations}$=*apriori(Data, Signs, Rules)*.

Data Mining є набором методик, який дає змогу визначити найсприйнятливіші для продукту, що просувається, або послуги категорії споживачів, виявити особливості найбільш успішних працівників, передбачити поведінкову модель споживачів тощо [10–12], тобто:

$$Md_{Data\ Mining} = \langle MD_{Supervised\ Learning}, MD_{Unsupervised\ Learning}, MD_{St}, MD_{Cb} \rangle, \quad (7)$$

де $MT_{Data\ Mining}$ – множина методів Data Mining для задачі класифікації;

$MD_{SLearning}$ – множина методів Data Mining навчання з учителем (Supervised Learning);

$MD_{ULearning}$ – множина методів навчання без учителя (Unsupervised Learning);

MDSt – статистичні методи Data Mining для аналізу BD;

MDCb – кібернетичні методи Data Mining для аналізу BD.

Інша класифікація методів Data Mining ґрунтується на різних підходах щодо навчання математичним моделям (рис. 4, 5) [14–18].

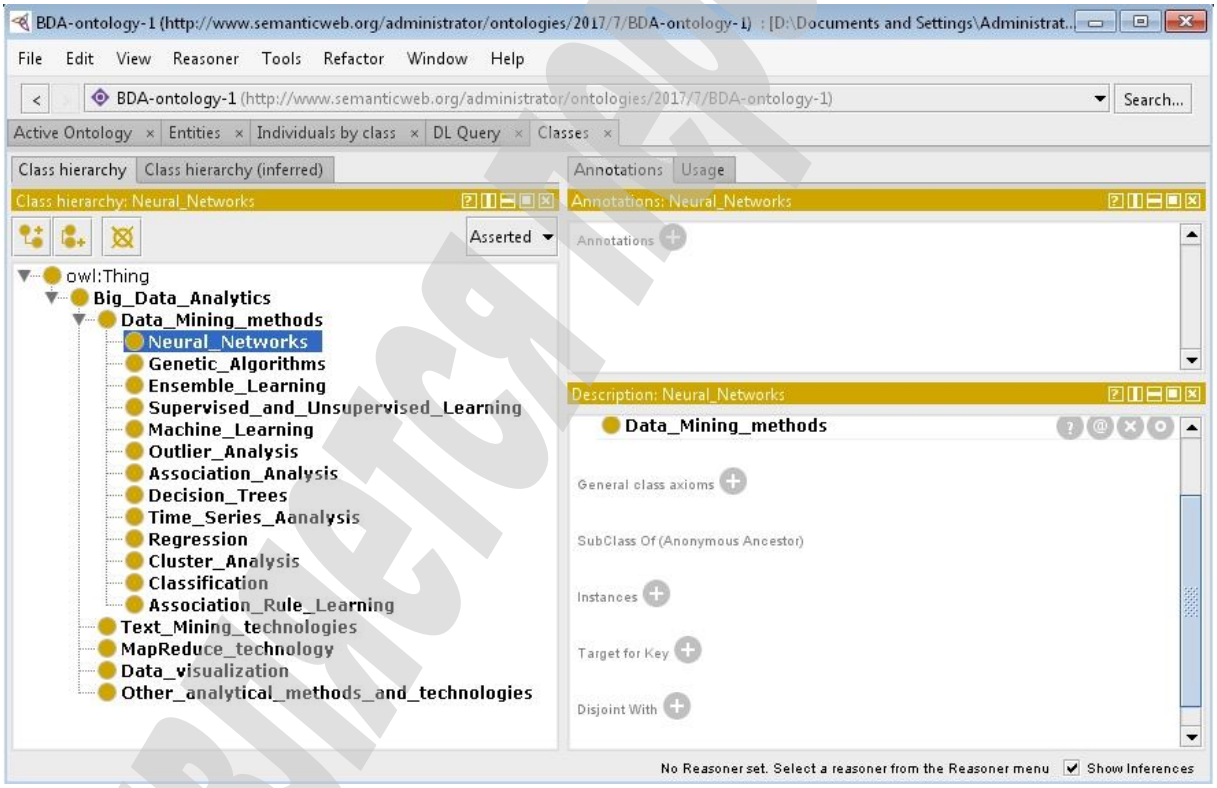

Рис. 4. Підкласи класу Data Mining Methods

Статистичні методи Data Mining містять: попередній аналіз природи статистичних даних, виявлення зв'язків і закономірностей, багатовимірний статистичний аналіз, динамічні моделі і прогноз на основі часових рядів:

$$MD_{St} = \langle MS_1, MS_2, MS_3, MS_4 \rangle, \quad (8)$$

де MS_1 – дескриптивний аналіз і опис початкових даних; MS_2 – аналіз зв'язків (кореляційний і регресійний, факторний, дисперсійний); MS_3 – багатовимірний статистичний аналіз (компонентний, дискримінантний, багатовимірний регресійний, канонічні кореляції); MS_4 – аналіз часових рядів (динамічні моделі і прогнозування).

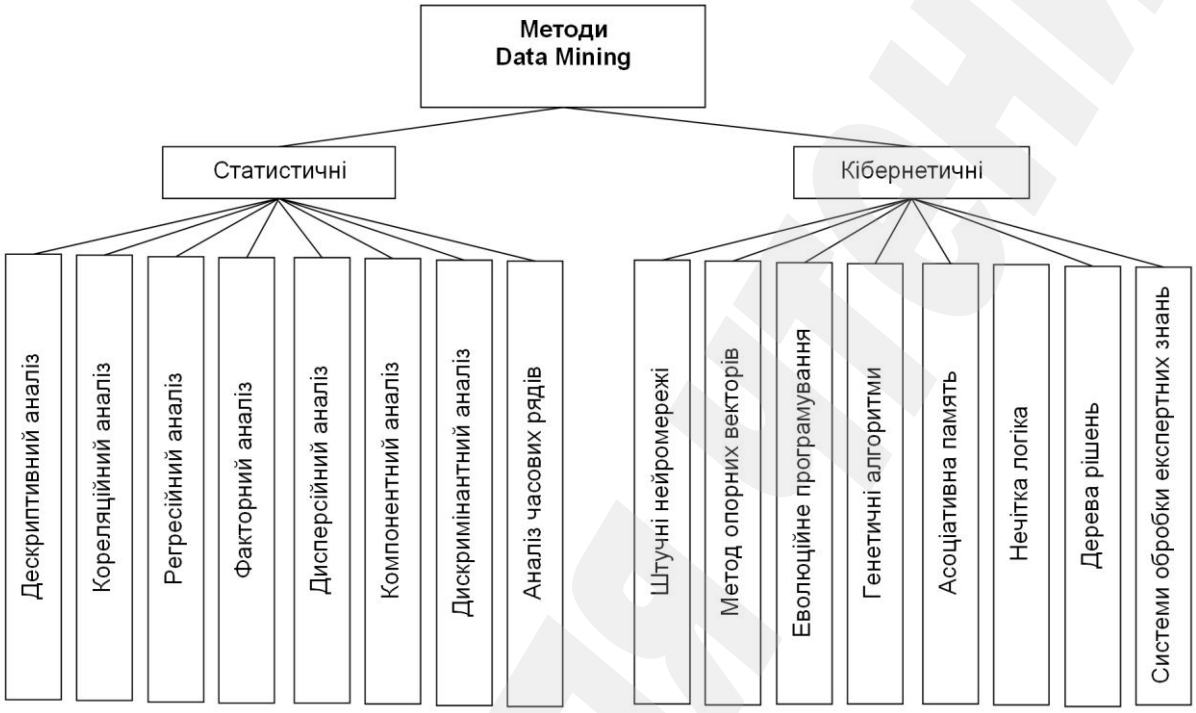

Рис. 5. Методи інтелектуального аналізу Big Data

Для вирішення завдання кластеризації на графах застосовують алгоритм Girvanand Newman, методу MLP (Markov Cluster Algorithm). Приклад – сегментування ринку. Регресійний аналіз використовують для визначення:

- впливу рівня задоволеності клієнтів на лояльність клієнтів;
- залежності кількості опор прийнятих викликів від прогнозу погоди, з огляду на попередній день;
- впливу сусідства і розміру на ціну лістингу будинків;
- сумісності у житті користувача через онлайн сайт знайомств тощо.

Аналіз часових рядів – відстеження ринку цінних паперів або захворюваності пацієнтів. Аналіз викидів застосовують для виявлення шахрайства, особистого маркетингу, медичного аналізу [19–23]. До кібернетичних методів DataMining належать такі методи [23–27]:

$$MD_{Cb}=\langle MC_1, MC_2, MC_3, MC_4, MC_5, MC_6, MC_7\rangle, \quad (9)$$

де MC_1 – еволюційне програмування; MC_2 – асоціативна пам'ять (пошук аналогів, прототипів); MC_3 – нечітка логіка; MC_4 – дерева рішень; MC_5 – системи опрацювання експертних знань; MC_6 – штучні нейронні мережі (розпізнавання, кластеризація, прогноз); MC_7 – генетичні алгоритми (оптимізація).

MC_6 – це клас моделей, що ґрунтуються на аналогії з роботою мозку людини та призначаються для розв'язання різноманітних задач аналізу даних після проходження етапу навчання на даних. MC_6 – це модель біологічних нейронних мереж мозку, в яких нейрони імітуються однотипними елементами (штучними нейронами). MC_6 застосовують для розв'язання таких задач:

- автоматизація процесів розпізнавання образів;
- прогнозування показників діяльності підприємства;
- медична діагностика; прогнозування;
- адаптивне управління;
- створення експертних систем;
- організація асоціативної пам'яті;
- оброблення аналогових та цифрових сигналів;
- синтез та ідентифікація електронних систем. За допомогою MC_6 можна, наприклад, передбачати обсяги продаж, показники фінансового ринку, розпізнавати сигнали, розробляти самонавчальні системи;

MC_7 навіяні природою еволюційних процесів, тобто за допомогою таких механізмів, як успадкування, мутації і природного добору. Ці механізми використовують для еволюціонування корисного вирішення проблем, які вимагають оптимізації. MC_7 використовують для розв'язку таких задач:

- формування розкладу лікарів для лікарні швидкої допомоги;
- створення комбінації оптимальних матеріалів та інженерних методів, необхідних для розробки економічних автомобілів;
- генерування штучно творчого контенту, такого як каламбури і жарти;
- прогнозування індексу фондового ринку за допомогою аналізу часових рядів. Для аналізу ринкового кошика застосовують аналіз прихованих закономірностей (Association Analysis). Статистичну класифікацію використовують для автоматичного призначення категорії документам; класифікації організмів на групи; розробленння профілів студентів, які приймають онлайн-курси; цілеспрямованого підбору працівників (focuse dhiring) тощо.

Ще одним цікавим напрямом в штучному інтелекті (ШІ) для аналізу BD є Machine Learning (ML). Це напрям в інформатиці (історично за ним закріпилася назва «штучний інтеллект»), який має на меті створення алгоритмів самонавчання на основі аналізу емпіричних даних. ML утворений як результат поділу науки про нейромережі на методи навчання мереж і види топологій архітектури мереж. Він також містить методи математичної статистики. Зазначені нижче способи ML ґрунтуються на нейромережах. Базові види нейромереж, а саме, перцептрон та багатошаровий перцептрон (та їхні модифікації) можуть навчатися як з учителем, без вчителя, з підкріпленням і активно. Однак, більшість статистичних методів і деякі нейромережі належать тільки до одного зі способів навчання:

$$M_{Machine\ Learning} = \langle ML_1, ML_2, ML_3, ML_4, ML_5, ML_6, ML_7, ML_8, ML_9, ML_{10}, fml, K_{Machine\ Learning} \rangle, \quad (10)$$

де ML_1 – методи навчання з учителем;

ML_2 – методи навчання без учителя;

ML_3 – методи навчання з підкріпленням;
ML_4 – методи активного навчання;
ML_5 – методи навчання з частковим залученням учителя;
ML_6 – методи трансдуктивного навчання;
ML_7 – методи багатозадачного навчання;
ML_8 – методи різноманітного навчання;
ML_9 – методи керованого і некерованого навчання;
ML_{10} – методи ансамблю навчання.

ML містить ПЗ, яке здатне видобувати знання з даних. Це дає ІС можливість вчитися, не будучи явно запрограмованими, та фокусується на прогнозуванні на основі відомих властивостей, видобутих з наборів навчальних даних. Сьогодні ML використовують:

- для розпізнавання спам і не спаму повідомлень е-пошти;
- для отримання знань про переваги користувача та надання рекомендацій, що ґрунтуються на даній інформації;
- для визначення кращого контенту для залучення потенційних клієнтів;
- для отримання ймовірності виграшу справи та встановлення юридичних норм пред'явлених рахунків.

ML_9 дають змогу виявити функціональні взаємозв'язки в аналізованих масивах даних. ML_{10} використовує множину предикативних моделей за рахунок чого підвищується якість зроблених прогнозів.

Методи графічного подання результатів аналізу BD $M_{Visualization}$ у вигляді діаграм або анімації значно спрощують інтерпретацію та полегшують розуміння отриманих результатів [27–31]. Новими прогресивними методами візуалізації є:

$$M_{Visualization} = \langle MV_1, MV_2, MV_3, MV_4, fmv, K_{Visualization} \rangle, \quad (11)$$

де MV_1 – хмара тегів; MV_2 – кластерграма; MV_3 – історичний потік; MV_4 – просторовий потік.

Кожному елементу в MV_1 присвоюють певний ваговий коефіцієнт, який корелює з розміром шрифту [32–36]. У разі аналізу тексту величина вагового коефіцієнта безпосередньо залежить від частоти вживання (цитування) певного слова або словосполучення. Дає змогу читачеві в стислі терміни отримати уявлення про ключові моменти скільки завгодно великого тексту або набору текстів. MV_2 показує як окремі елементи множини даних співвідносяться з кластерами в міру зміни їхньої кількості. Вибір оптимальної кількості кластерів – важлива складова кластерного аналізу. MV_3 допомагає стежити за еволюцією документа, над створенням якого працює одночасно велика кількість авторів. По горизонтальній осі відкладається час, по вертикальній – внесок кожного з співавторів, тобто обсяг введеного тексту. Кожному унікальному автору присвоюється певний колір на діаграмі. MV_4 дає змогу відстежувати просторовий розподіл інформації. Чим яскравіше лінія – тим більше даних передається за одиницю часу [36–37].

Підґрунтям $T_{Text\ Mining}$ є статистичний та лінгвістичний аналіз, методи штучного інтелекту. Дана технологія застосовується для проведення аналізу, забезпечення навігації та пошуку в неструктурованих текстах [38–42]. Застосування ІТ класу $T_{Text\ Mining}$ дає змогу користувачам набувати нових знань. Це набір ме-

тодів, які призначені для видобування відомостей з текстів на основі сучасних ІТ, що дає змогу виявити закономірності, та забезпечити отримання користувачами корисних даних та нових знань (рис. 6):

$$T_{Text\ Mining} = \langle MT_1, MT_2, MT_3, MT_4, MT_5, MT_6, MT_7, MT_8, MT_9, MT_{10}, MT_{11}, fmt, K_{Text\ Mining} \rangle, \quad (12)$$

де MT_1 – класифікація (*classification*);

MT_2 – кластеризація (*clustering*);

MT_3 – побудова семантичних мереж або аналіз зв'язків (*Relationship, Event andFact Extraction*);

MT_4 – здобуття феноменів, фактів, понять (feature extraction);

MT_5 – автоматичне реферування, створення анотацій (summarization);

MT_6 – відповідь на запити (question answering);

MT_7 – тематичне індексування (thematic indexing);

MT_8 – пошук за ключовими словами (keyword searching);

MT_9 – засоби підтримки та створення таксономії (oftaxonomies);

MT_{10} – засоби підтримки та створення тезаурусів (thesauri);

MT_{11} – методи та засоби контент-аналізу (Content Analysis).

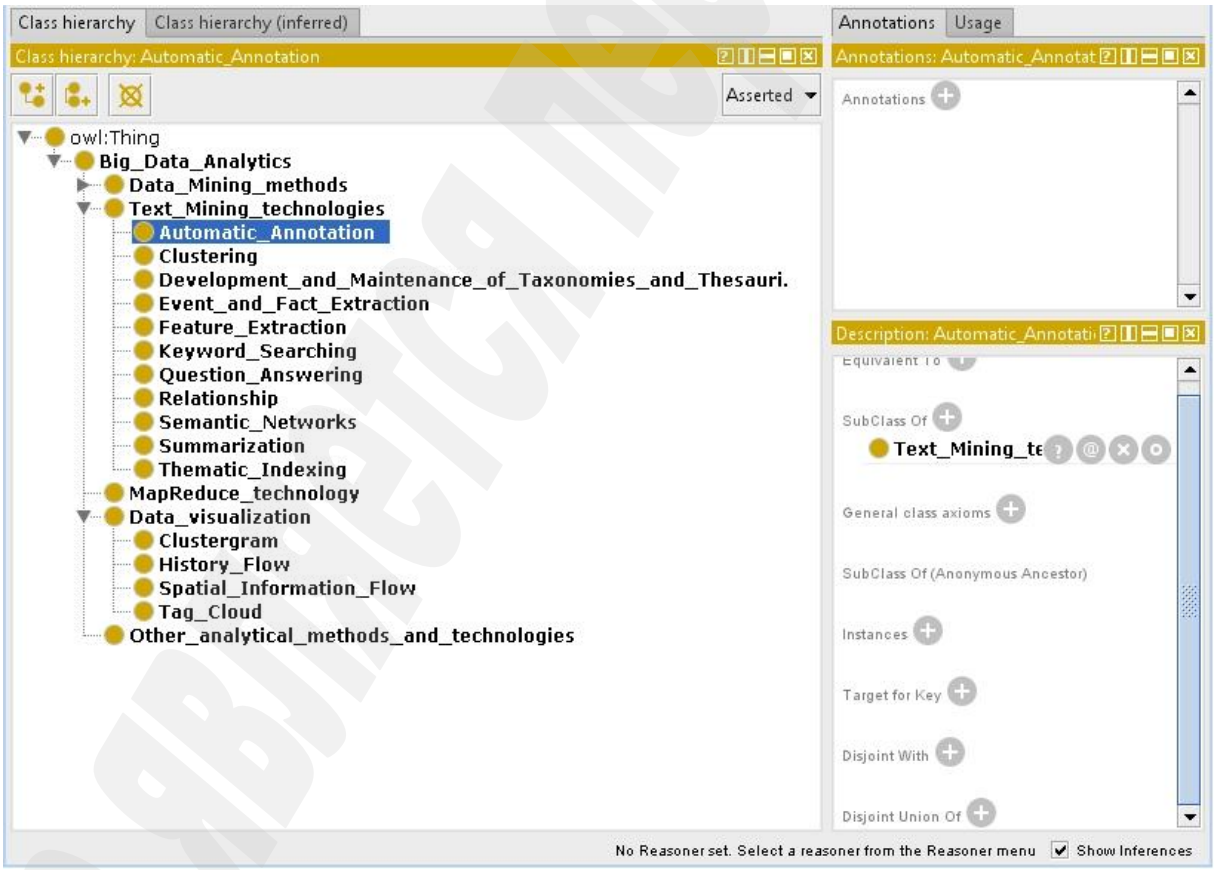

Рис. 6. Підкласи класу Text Mining Technologies

$T_{Text\ Mining}$, як і більшість когнітивних технологій, – це алгоритмічне виявлення раніше не відомих зв'язків та кореляцій в уже наявних текстових даних. У $T_{Text\ Mining}$ широко використовують методологію і підходи технології аналізу

видобування даних, наприклад, MT_1 чи MT_2. $T_{Text\ Mining}$ мають нові можливості: автоматичне реферування текстів та виявлення феноменів, тобто понять і фактів. Важливим завданням $T_{Text\ Mining}$ є видобування з тексту його характерних елементів або властивостей, які можуть використовуватися як метадані документа, ключові слова, анотації. Іншим важливим завданням є встановлення приналежності документу до певних категорій зі заданої схеми їхньої систематизації. $T_{Text\ Mining}$ забезпечують новий рівень семантичного пошуку документів. Можливості $T_{Text\ Mining}$ застосовують для розв'язання задачі виявлення шаблонів в тексті, автоматичного «виштовхування» або розподілу даних за профілями, створення оглядів документів.

Це інструмент, який дає можливість аналізувати BD у пошуках тенденцій, шаблонів та взаємозв'язків, здатних допомогти при ухваленні стратегічних рішень. Основна мета $T_{Text\ Mining}$ надати аналітику можливість працювати з BD за рахунок автоматизації процесу здобуття потрібних даних. Як приклад ефективного застосовування технологій $T_{TextMining}$ є MT_{11}, який характеризується об'єктивністю висновків та строгістю процедури. Його основою є квантифікація тексту з подальшою інтерпретацією результатів. Предметом MT_{11} можуть бути як проблеми соціальної дійсності, які висловлюються чи навпаки приховуються у документах, так і внутрішні закономірності самого об'єкту дослідження [38]. Популярність MT_{11} ґрунтується на тому, що цей метод дає змогу виміряти людську поведінку (якщо вважати, що вербальна поведінка є її формою). На відміну від опитувань, контент-аналіз вимірює не те, що люди говорять, що зробили чи зроблять, а те що вони справді зробили.

Опишемо декілька технологій і дисциплін дослідження даних з погляду технології BD для T_{other} (рис. 7) [14–19]:

$$T_{other} = \langle MO_1, MO_2, MO_3, MO_4, MO_5, MO_6, MO_7, MO_8, MO_9, MO_{10}, MO_{11}, MO_{12}, MO_{13}, fmo, K_{other} \rangle, \quad (13)$$

де MO_1 – методи A/B тестування (A/B testing, Splittesting);
MO_2 – методи опрацювання природної мови (Natural Language Processing, NLP);
MO_3 – методи аналізу настроїв (Sentiment Analysis);
MO_4 – методи мережевого аналізу (Network Analysis);
MO_5 – методи оптимізації (Optimization);
MO_6 – методи розпізнавання образів (Pattern Recognition);
MO_7 – методи прогнозного моделювання (Predictive Modeling);
MO_8 – методи опрацювання сигналів (Signal Processing);
MO_9 – методи просторового аналізу (Spatial Analysis);
MO_{10} – методи статистики (Statistics);
MO_{11} – методи моделювання (Simulation);
MO_{12} – методи краудсорсінгу (Crowdsourcing);
MO_{13} – методи злиття та інтеграція даних(Data Fusion and Data Integration).

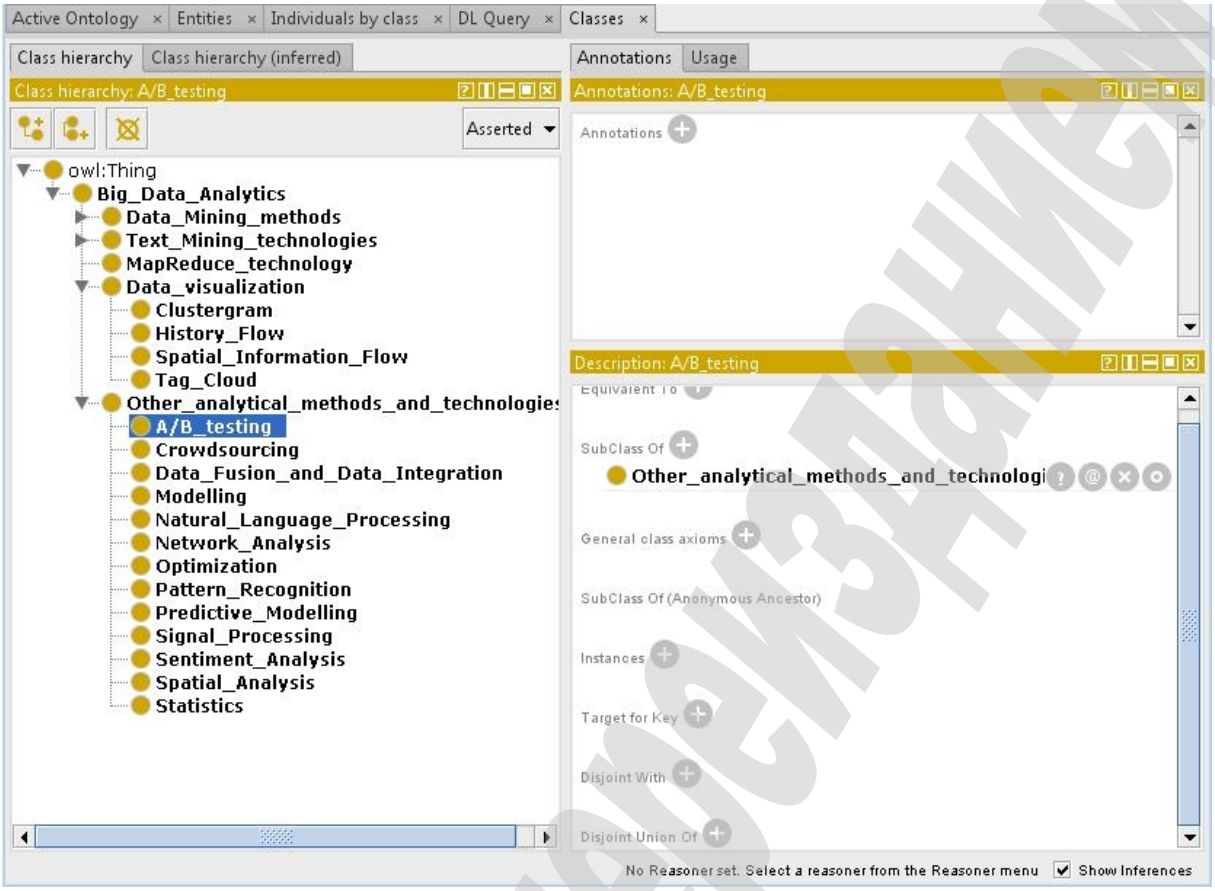

Рис. 7. Підкласи класу Other technologies and research techniques

MO$_1$ використовують при оптимізації Web-сторінок відповідно до заданої мети. В основі *MO*$_3$ лежать *MO*$_2$. Вони дають змогу виокремити із загального інформаційного потоку повідомлення, пов'язані зі зацікавленим предметом (наприклад, споживчим продуктом). Далі оцінити полярність судження (позитивне чи негативне), ступінь емоційності тощо. *MO*$_3$ допомагає дослідникам визначити настрої спікерів або авторів по відношенню до теми. Аналіз настроїв використовується, щоб допомогти: поліпшити якість обслуговування в готельній мережі, аналізуючи коментарі гостей; налаштувати стимули і послуги для вирішення того, що клієнти дійсно просять; визначити, які споживачі дійсно є під впливом соціальних медіа. *MO*$_4$ є методикою аналізу зв'язків між вузлами в мережах. Стосовно до соціальних мереж дає змогу аналізувати взаємозв'язок між окремими користувачами, компаніями, спільнотами тощо. *MO*$_5$ призначений для редизайну складних систем і процесів для поліпшення одного або декількох показників. Допомагає в прийнятті стратегічних рішень, наприклад, складу виведеної на ринок продуктової лінійки, проведенні інвестиційного аналізу тощо. *MO*$_7$ дають змогу створити математичну модель наперед заданого ймовірного сценарію розвитку подій. Наприклад, аналіз бази даних CRM-системи на предмет можливих умов, які підштовхнуть абонентів змінити провайдера. *MO*$_{12}$ – категоризація та збагачення даних силами широкого, невизначеного кола осіб, з метою використання їхніх творчих здібностей, знань і досвіду зі застосуванням інформаційно-комунікаційних технологій. *MO*$_{13}$ дає змогу аналізувати коментарі користувачів соціальних мереж і зіставляти з результатами продажів в режимі реального часу.

T$_{MapReduce}$ – модель розподілених обчислень, представлена компанією Google, яка використовується для паралельних обчислень над дуже великими (кілька петабайт)

наборами даних в комп'ютерних кластерах [42–47]. З погляду реалізації, аналітична платформа для роботи з BD має вміти використовувати нові $T_{MapReduce}$. На практиці, аналіз BD рідко полягає в тому, щоб обчислити статистичні підсумки за всіма даними. Замість цього значимість BD полягає в можливості поділу даних на мікросегменти і за допомогою методів інтелектуального аналізу та прогностичного моделювання побудувати велику кількість моделей для невеликих груп спостережень. Є безліч інструментів для проведення такого агрегування даних в розподіленій файловій системі, що дає змогу легко здійснювати даний аналітичний процес.

6. Результати досліджень

Приведений опис методів і технологій аналізу BD дає змогу побудувати онтологію відповідно до підходу METHONTOLOGY [46–50], який відображає процес ітеративного проектування. За методологією METHONTOLOGY глосарій термінів містить всі терміни (концепти і їхні екземпляри, атрибути, дії), важливі для аналізу BD, і їхні природно-мовні описи. Глосарій термінів онтології аналізу BD містить означені вище терміни, які можна семантично розбити на три групи:

1) структура завдання (групи технологій аналітики, зв'язки);

2) дані, що наповнюють задачу (методи, що застосовують для кожної групи);

3) результати обчислень (рекомендації щодо використання BD для підвищення ефективності ухвалення рішень) [50–55].

Розроблена засобами Protégé-OWL онтологія аналізу BD подана на рис. 8.

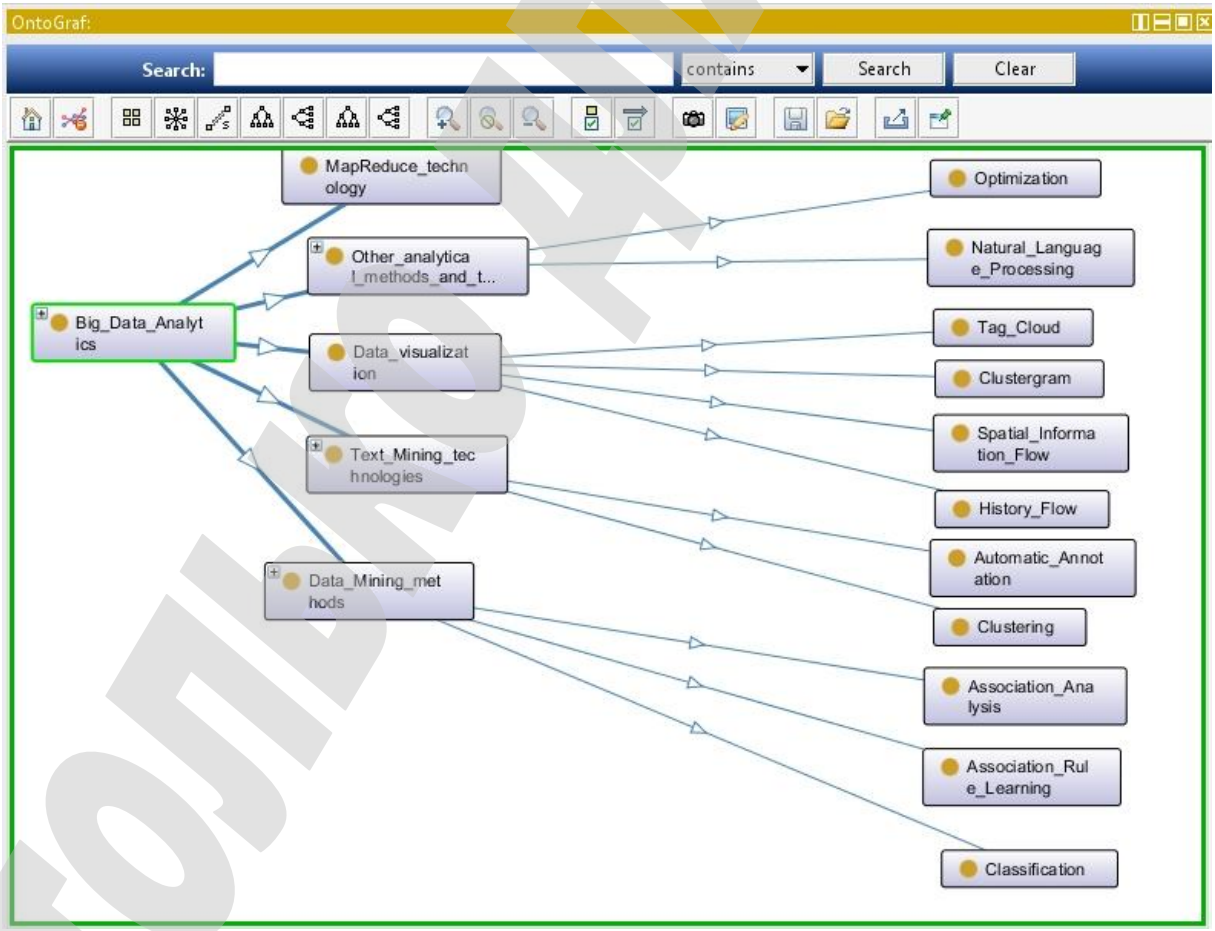

Рис. 8. Структура онтології для аналізу Big Data у вигляді графу

Кожна BD володіє набором параметрів та критеріїв, які визначають методики та технології аналізу K_U. Саме призначення BD, її структура та наповнення визначають методики та технології подальшого аналізу аналізу.

Завдяки розробленій онтології БЗ аналізу BD за допомогою Protégé 3.4.7 та побудованих в них множини правил RABD можна скоротити процес вибору методотик та технологій для подальшого аналізу та полегшити автоматизацію самого процесу аналізу обраної BD. Так, $K_U \bigcup K_{BD}$ при $K_U \subset K_{BD}$ дозволить сформувати нову множину K_U для визначення A_{BD}:

$$K_U = K'_{Data\ Mining} \bigcup K'_{MachineLearning} \bigcup K'_{Visualization} \bigcup$$
$$\bigcup K'_{Text\ Mining} \bigcup K'_{Map\,\mathrm{Re}duce} \bigcup K'_{other},$$

де $K'_{Data\ Mining} \subseteq K_{Data\ Mining}$, $K'_{MachineLearning} \subseteq K_{MachineLearning}$, $K'_{Visualization} \subseteq K_{Visualization}$, $K'_{Text\ Mining} \subseteq K_{Text\ Mining}$, $K'_{Map\,\mathrm{Re}duce} \subseteq K_{Map\,\mathrm{Re}duce}$, $K'_{other} \subseteq K_{other}$.

Тоді отримаємо нове значення:

$$A'_{BD} = \langle M'_{Data\ Mining}, M'_{Machine\ Learning}, M'_{Visualization},$$
$$T'_{Text\ Mining}, T'_{MapReduce}, T'_{other}, K_U \rangle.$$

Наприклад, для аналізу BD соціальних мереж [55] критеріями та параметрами є склад профілів користувачів (тексти, гіпертексти, вік, відношення, стать, країна, кількість друзів, пости, активність тощо). Також критерієм аналізу є взаємодії з іншими користувачами як цієї соціальної мережі, так і поза нею. Причому користувачами є не лише конкретні особистості, але можуть бути інформаційні ресурси та агенти. Застосувавши розроблену онтологію БЗ (рис. 9) для BD соціальної мережі, отримаємо:

$$M'_{Data\ Mining} = fdm(Tk'_{Data\ Mining}, Md'_{Data\ Mining}, K'_{Data\ Mining}),$$

де $Tk'_{Data\ Mining} = \langle T_{Classification}, T_{Clustering}, T_{Sequence}, T_{LinkAnalysis}, T_{Graph\ Mining}, T_{Summarization} \rangle$,
$MT'_{Data\ Mining} = \langle M_{Bayesian\ Networks}, M_{Tree}, M_{Neural\ Networks} \rangle$,
$Md'_{Data\ Mining} = \langle MD_{St}, MD_{Cb} \rangle$,
$MD'_{St} = \langle MS_3, MS_4 \rangle$, $MD_{Cb} = \langle MC_3, MC_4, MC_6 \rangle$,
$M'_{Machine\ Learning} = fml\ (ML_5, K'_{Machine\ Learning})$,
$M'_{Visualization} = fmv\ (MV_1, MV_2, MV_3, MV_4, K'_{Visualization})$,
$T'_{Text\ Mining} = fmt\ (MT_1, MT_2, MT_3, MT_6, MT_7,$
$MT_8, MT_9, MT_{10}, MT_{11}, K'_{Text\ Mining})$,
$T'_{other} = fmo\ (MO_1, MO_2, MO_3, MO_4, MO_9, MO_{10}, MO_{12}, MO_{13}, K'_{other})$.

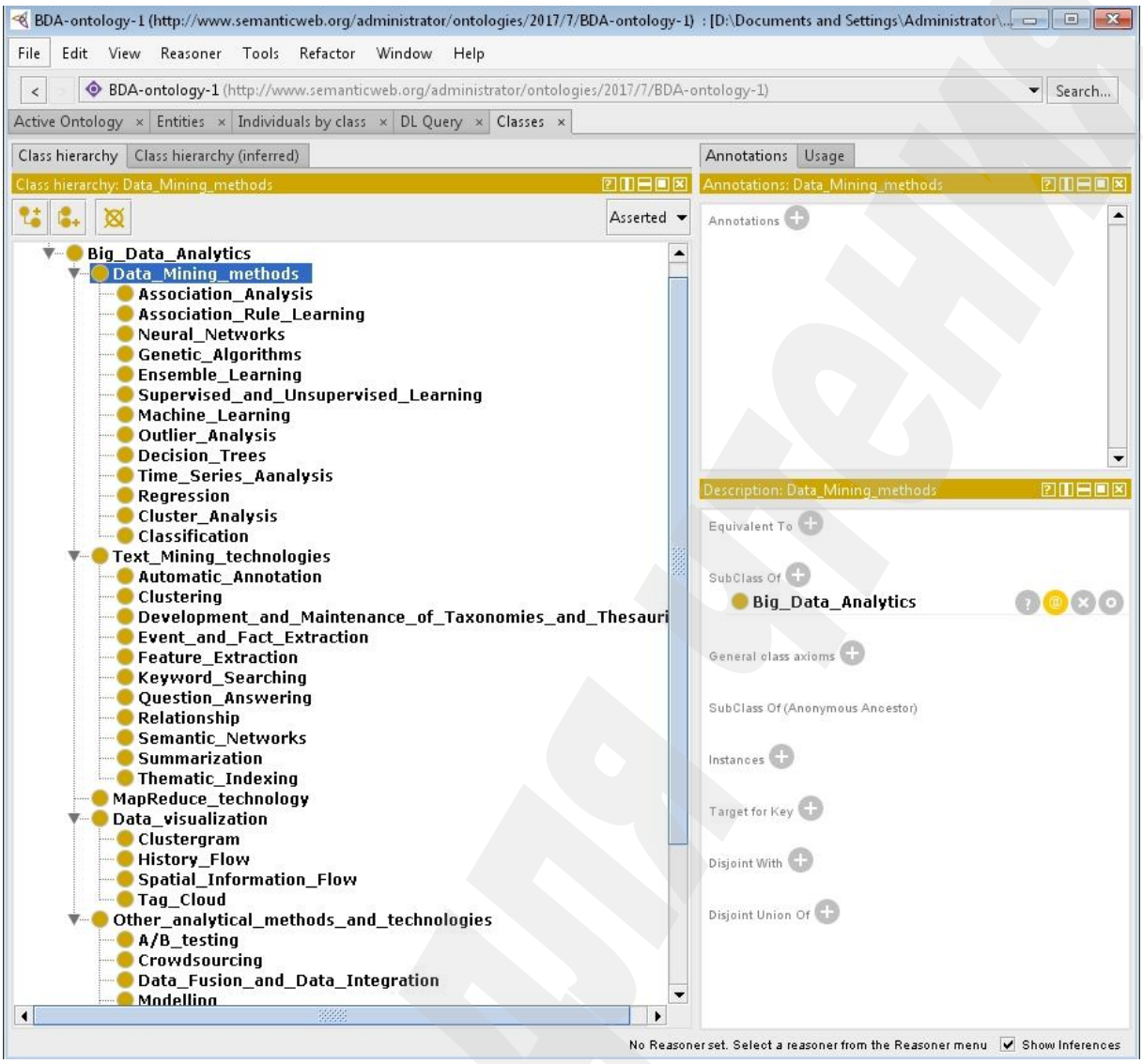

Рис. 9. Ієрархія класів онтології для аналізу Big Data

Оптимальне визначення всіх критеріїв та параметрів аналізу соціальної мережі дозволить ефективно застосувати результати аналізу для ідентифікації напрямів взаємовідносин між людьми в багатьох областях та в комерційній діяльності. Вузли є користувачами мережі, в той час як зв'язки є відносинами між ними. Аналіз соціальних мереж використовується для розв'язку задач такого типу:

- як люди з різних популяцій утворюють зв'язки зі сторонніми;
- знайти значення або ступінь впливу конкретного індивіда в групі;
- знайти мінімальну кількість прямих зв'язків, потрібних для підключення двох осіб;
- зрозуміти соціальну структуру клієнтської бази. Можна прослідкувати популяризацію тематики/товару в залежності від віку, статті, країни проживання, статусу та рівня освіти множини користувачів конкретної соціальної мережі.

7. SWOT-аналіз результатів досліджень

Strengths. Наявність онтологічної БЗ як фундаментального класифікатора для вибору оптимального алгоритму аналізу BD відповідної до її стуркутри та ПЗ. Класифікатор дозволяє визначити множину правил аналізу Big Data RABD

з метою її використання для опрацювання та аналізу конкретної BD на основі її параметрів та критеріїв.

Weaknesses. Неможливість формування множини правил аналізу Big Data RABD при відсутності декількох критеріїв та параметрів конкретної BD. Неточне визначення критерію/параметру конкретної BD призведе до формування неефективного алгоритму аналізу BD та збільшить трудомісткість обчислення.

Opportunities. Перспективи подальших досліджень полягатимуть у дослідженні методів, моделей та інструментів для удосконалення онтології аналітики BD та ефективнішої підтримки розроблення структурних елементів моделі системи підтримки прийняття рішень з керування BD.

Threats. Негативна дія на об'єкт дослідження зовнішніх чинників флотування множини критеріїв та параметрів аналізу BD. Відсутність в світі аналогів даного об'єкта дослідження та проведених масштабних експериментів на основі розробленої моделі не дає чітких напрямів подальших досліджень.

8. Висновки

1. Досліджено особливості класифікації методів і технологій аналітики Big Data з врахуванням означення та особливості застосування відповідних IT. Особливості застосування методів Data Mining, технологій Text Mining, технології MapReduce, візуалізації даних, інших технологій та методик аналізу BD дало змогу побудувати онтологію відповідно до підходу METHONTOLOGY. Він відображає процес ітеративного проектування та дозволяє побудувати глосарій термінів, важливих для аналізу BD, і їхні природно-мовні описи. Розроблений глосарій термінів онтології аналізу BD містить необхідні терміни типу структур завдання, дані задачі та результати обчислень. Чим повніший глосарій, тим ефективніший отриманий результат у вигляді алгоритму аналізу BD.
2. Розроблено формальну модель аналізу BD. На вході системи є методи та IT аналізу BD. На виході системи є онтологічна модель правил аналізу BD.
3. Розроблено онтологічну БЗ аналізу BD. Таксономія понять онтології задає методику аналізу BD. Оптимальне визначення множини відношень між цими поняттями та множини правил аналізу BD, формалізованих за допомогою дескриптивної логіки DL, дозволяє ефективно опрацьовувати BD.
4. Побудовані правила аналізу Big Data RABD. Кожна BD володіє набором параметрів та критеріїв, які визначають методики та технології аналізу. Саме призначення BD, її структура та наповнення визначають методики та технології подальшого аналізу аналізу. Завдяки розробленій онтології БЗ аналізу BD за допомогою Protégé 3.4.7 та побудованих в них множини правил RABD скорочується процес вибору методотик та технологій для подальшого аналізу та автоматизується процес аналізу обраної BD.

Подяка

Роботу виконано в рамках спільних наукових досліджень кафедри інформаційних систем та мереж (ІСМ) Національного університету «Львівська політехніка» (Україна) на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів

баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ICM Національного університету «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

Літерарура

1. Mayer-Schonberger V., Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. John Murray Publishers, 2013. 256 p.
2. Fekete J. D. Big Data Visual Analytics. 2016. URL: http://www.aviz.fr/wiki/uploads/TeachingVA2016/Lectur-BigDataVA.pdf (Last accessed: 18.09.2017).
3. Raghupathi W., Raghupathi V. Big data analytics in healthcare: promise and potential // Health Information Science and Systems. 2014. Vol. 2, No. 1. doi:10.1186/2047-2501-2-3
4. Hong S. H., Ma K. L., Koyamada K. Big Data Visual Analytics. NII Shonan Meeting Report No. 2015-147. Tokyo, 2017. URL: https://pdfs.semanticscholar.org/45ec/4934ee034a5839f4e657089ac865f0baa8ff.pdf (Last accessed: 18.09.2017).
5. MAD Skills: New Analysis Practices for Big Data / Cohen J. et al. // Proceedings of the VLDB Endowment. 2009. Vol. 2, No. 2. P. 1481–1492. doi:10.14778/1687553.1687576
6. History and evolution of big data analytics. URL: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html (Last accessed: 18.09.2017).
7. Mitchell R. L. 8 big trends in big data analytics. URL: http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html (Last accessed: 18.09.2017).
8. Big Data. URL: http://tadviser.ru/a/125096 (Last accessed: 18.09.2017).
9. Inmon W. H. Big Data – getting it right: A checklist to evaluate your environment. Forest Rim Technology LLC. 2014. URL: http://dssresources.com/papers/features/inmon/inmon01162014.htm (Last accessed: 18.09.2017).
10. Analysis of data and processes / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2009. 512 p.
11. Paklin N. B., Oreshkov V. I. Business analysis: from data to knowledge. Saint Petersburg: Piter, 2009. 624 p.
12. Duke V., Samoylenko A. Data Mining: training course. Saint Petersburg: Piter, 2001. 368 p.
13. Manyika J. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. 156 p.
14. Zhuravlev J. I., Ryazanov V. V., Senko O. V. Recognition. Mathematical methods. Software system. Practical applications. Moscow: Phasis, 2006. 176 p.
15. Zinovev A. Y. Visualization of multidimensional data. Krasnoyarsk: Publisher Krasnoyarsk State Technical University, 2000. 180 p.
16. Chubukova I. A. Data Mining: A Tutorial. Moscow: Internet University of Information Technologies, BINOM, 2006. 382 p.
17. Sitnik V. F., Krasnyuk M. T. Data Mining. Kyiv: KNEU, 2007. 376 p.

18. Witten I. H., Frank E., Hall M. A. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann, 2011. 664 p. doi:10.1016/c2009-0-19715-5
19. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. John Wiley & Sons Ltd, 2015. 256 p.
20. Einav L., Levin J. The Data Revolution and Economic Analysis. 2014. URL: http://www.nber.org/chapters/c12942.pdf (Last accessed: 18.09.2017).
21. Vanyashin A., Klimentov A., Korenkov V. PANDA follows the large data // Supercomputers. 2013. Vol. 3, No. 11. P. 56–61.
22. Serov D. Analytics of «big data» – new perspectives. URL: http://www.storagenews.ru/49/EMC_BigData_49.pdf (Last accessed: 18.09.2017).
23. Links that speak: The global language network and its association with global fame / Ronen S. et al. // Proceedings of the National Academy of Sciences. 2014. Vol. 111, No. 52. P. 5616–5622. doi:10.1073/pnas.1410931111
24. Aflalo Y., Kimmel R. Spectral multidimensional scaling // Proceedings of the National Academy of Sciences. 2013. Vol. 110, No. 45. P. 18052–18057. doi:10.1073/pnas.1308708110
25. Gadepally V., Kepner J. Big data dimensional analysis // 2014 IEEE High Performance Extreme Computing Conference (HPEC). 2014. doi:10.1109/hpec.2014.7040944
26. Analyzing Big Data with Dynamic Quantum Clustering / Weinstein M. et al. URL: https://arxiv.org/ftp/arxiv/papers/1310/1310.2700.pdf (Last accessed: 18.09.2017).
27. Paklin N. B., Oreshkov V. I. Business Intelligence: from data to knowledge. Saint Petersburg: Piter, 2013. 702 p.
28. Zelazny D. Speak in the language of diagrams: manual on visual communications for managers. Moscow: Institute for Comprehensive Strategic Studies, 2004. 220 p.
29. Roem D. The practice of visual thinking. An original method for solving complex problems. Moscow: Mann, Ivanov and Ferber, 2014. 396 p.
30. Russom P. Big data analytics. 2011. URL: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf (Last accessed: 18.09.2017).
31. Yau N. The art of visualization in business. How to present complex information with simple images. Moscow: Mann, Ivanov and Ferber, 2013. 352 p.
32. Iliinsky N., Steele J. Designing Data Visualizations. Sebastopol: O'Reilly, 2011. 110 p.
33. Krum R. Cool infographics: effective communication with data visualization and design. Indianapolis: Wiley, 2014. 348 p.
34. Tukey J. Analysis of Observation Results: Exploratory Analysis. Moscow: Mir, 1981. 693 p.
35. Alper C., Brown K., Wagner G. R. New Software for Visualizing the Past, Present and Future. 2006. URL: http://dssresources.com/papers/features/alperbrown&wagner/alperbrown&wagner09212006.html (Last accessed: 18.09.2017).
36. Analysis of data and processes / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2009. 512 p.
37. Text Mining. URL: http://statsoft.ru/home/textbook/modules/sttextmin.html#index (Last accessed: 18.09.2017).
38. Lande D., Berezin B., Pavlenko O. Postroenie modeli informatsionnogo servisa na baze natsional'nogo segmenta Internet // Informatsionnye tehnologii i bezopasnost'. Materialy XVI Mezhdunarodnoi nauchno-prakticheskoi konferentsii ITB-2016. Kyiv: IPRI NAN Ukrainy, 2017. P. 48–57. URL: http://dwl.kiev.ua/art/itb2016/i4/i4.pdf (Last accessed: 18.09.2017).

39. Data Analysis Technologies. Data Mining, Visual Mining, Text Mining, OLAP / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2007. 384 p.

40. Linyuchev P. Text Mining: modern technologies on information mines // PC Week/RE. 2007. Vol. 6 (564). URL: https://www.pcweek.ru/idea/article/detail.php?ID=82081 (Last accessed: 18.09.2017).

41. Pleskach V. L., Zatonatskaya T. G. Information systems and technologies at enterprises. Kyiv: Znannya, 2011. 718 p.

42. MapReduce and Parallel DBMSs: Friends or Foes? / Stonebraker M. et al. // Communications of the ACM. 2010. Vol. 53, No. 1. P. 64. doi:10.1145/1629175.1629197

43. Berezin A. Map-Reduce on the example of MongoDB. 2013. URL: https://habrahabr.ru/post/184130/ (Last accessed: 18.09.2017).

44. Lebedenko E. Google MapReduce technology: divide and conquer. Kompiuterra, 2013. URL: http://www.computerra.ru/82659/mapreduce/ (Last accessed: 18.09.2017).

45. A comparison of approaches to large-scale data analysis / Pavlo A. et al. // Proceedings of the 35th SIGMOD International Conference on Management of Data – SIGMOD '09. 2009. doi:10.1145/1559845.1559865

46. Big Data from A to Ya. Part 1: Principles of working with large data, the MapReduce paradigm. 2015. URL: https://habrahabr.ru/company/dca/blog/267361/ (Last accessed: 18.09.2017).

47. Big Data from A to Ya. Part 3: Methods and strategies for developing MapReduce applications. 2015. URL: https://habrahabr.ru/company/dca/blog/270453/ (Last accessed: 18.09.2017).

48. Gavrilova T. A., Khoroshevsky V. F. Intelligent Systems Knowledge Base. Saint Petersburg: Piter, 2000. 384 p.

49. Classification Methods of Text Documents Using Ontology Based Approach / Lytvyn V. et al. // Advances in Intelligent Systems and Computing. Springer, 2016. P. 229–240. doi:10.1007/978-3-319-45991-2_15

50. Bisikalo O. V., Vysotska V. A. Identifying keywords on the basis of content monitoring method in Ukrainian texts // Radio Electronics, Computer Science, Control. 2016. Vol. 1, No. 36. P. 74–83. doi:10.15588/1607-3274-2016-1-9

51. Bisikalo O. V., Vysotska V. A. Sentence syntactic analysis application to keywords identification Ukrainian texts // Radio Electronics, Computer Science, Control. 2016. Vol. 3, No. 38. P. 54–65. doi:10.15588/1607-3274-2016-3-7

52. Lytvyn V., Bobyk I., Vysotska V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic // Radio Electronics, Computer Science, Control. 2016. Vol. 4, No. 39. P. 54–67. doi:10.15588/1607-3274-2016-4-10

53. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource management based on fuzzy logic // Radio Electronics, Computer Science, Control. 2015. Vol. 3, No. 34. P. 71–79. doi:10.15588/1607-3274-2015-3-9

54. Matches prognostication features and perspectives in cybersport / Korobchynskyi M. et al. // Radio Electronics, Computer Science, Control. 2017. Vol. 3, No. 42. P. 95–105. doi:10.15588/1607-3274-2017-3-11

55. Wolfram S. Data Science of the Facebook World. 2013. URL: http://blog.wolfram.com/2013/04/24/data-science-of-the-facebook-world/ (Last accessed: 18.09.2017).