УДК 004.89:002.53: 621.921.34 DOI: 10.33839/2708-731X-24-1-335-342

М.В. Дубенко; **В.М. Кулаківський**, канд. техн. наук

Інститут надтвердих матеріалів ім. В.М. Бакуля НАН України, вул. Автозаводська 2, 04074, м. Київ, e-mail: dubenko.marija@ism.kiev.ua

ОСНОВНІ ПРИНЦИПИ *DATA MINING*

У статті подано визначення терміну Data Mining, розглянуто його методи та процеси; наведено варіанти вирішення проблем, що виникають при аналізуванні наукометричних баз даних; розглянуто декілька структур сховищ даних; розглянуто метод аналізу OLAP та його можливості, проведено порівняння OLAP та Data Mining, наведено сфери застосування процесу Data Mining. Аналіз цієї інформації показав, що Data Mining – необхідний засіб для знаходження нових знань в сфері матеріалознавства і надтвердих матеріалів.

Ключові слова: *Data Mining, інтелектуальний аналіз даних, методи Data Mining, етапи Data Mining, Online Analytical Processing , OLAP.*

Ісаак Ньютон казав: «Якщо я бачив далі інших, то тому, що стояв на плечах гігантів». Це означає, що його винаходи були б неможливі без попереднього доробку великих вчених. Тільки тоді цей доробок було дуже важко отримати – потрібно було віднайти цю інформацію по частинкам у сотнях, чи навіть тисячах книг. Зараз цей процес можна зробити в тисячі разів швидшим і результативним завдяки наукометричним базам даних і різним методам їх аналізу, зокрема *Data Mining*.

Більше 10 років тому в Україні почали функціонувати світовий центр даних «Геоінформатика і сталий розвиток» і національна *Grid*-інфраструктура (академічний і освітянський сегменти), тому вітчизняні вчені і фахівці зараз мають у своєму розпорядженні великі обсяги даних з різних галузей, що обробляються в об'єднаній мережі кластерів країни. Розвиток методів запису і зберігання даних викликав бурхливе зростання об'ємів збираної і аналізованої інформації. Об'єми даних настільки значні, що людина просто не спроможна проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна – адже в цих «сирих даних» закладено знання, які можуть бути використані при ухваленні рішень. Наша єдина надія зрозуміти і знайти щось корисне в цьому океані інформації – широке застосування методів Data Mining. [1].

Рис. 1. Data mining – мультидисциплінарна область

Data Mining – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних для інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності. [2].

Технологія *Data Mining* вивчає процес знаходження нових, дійсних і потенційно корисних знань в базах даних. *Data Mining* лежить на перетині кількох наук, головні з яких – це системи баз даних, статистика, штучний інтелект та ін. (рис. 1).

Оскільки технологія *Data mining* – мультидисциплінарна

область, то для розробки програмного забезпечення, що включає інтелектуальний аналіз даних, необхідно задіяти фахівців з різних галузей, а також забезпечити їх високоякісну взаємодію. Неможливо видобувати корисну інформацію без розуміння сутності даних. Використання *Data mining* має бути нерозривно пов'язаним із підвищенням кваліфікації користувача. Більшість інструментів інтелектуального аналізу даних ґрунтується на двох технологіях: машинне навчання (machine learning) і візуалізація (візуальне подання інформації). Ці дві технології поєднують у собі байєсівські мережі (БМ). [3] Це відносно молодий напрям розвитку науки, що з'явився на стику теорії ймовірностей і теорії графів; він представляє набір випадкових змінних та їхніх умовних залежностей за допомогою орієнтованого ациклічного графу (ОАГ, англ. *directed acyclic graph, DAG*) [4].

Процес *Data Mining* є нескінченним циклом послідовних кроків, що допомагають ідентифікувати, вирішити і визначити нову задачу. Існує 4 етапи видобутку даних (рис. 2):

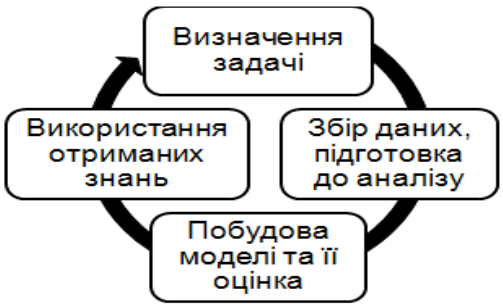


Рис. 2. Процес Data Mining

- визначення задачі (чітке визначення мети і вимог);
- збір даних і підготовка до аналізу (пошук і збір даних; визначення релевантності зібраних даних; видалення непотрібних даних; ідентифікація закономірностей і патернів; побудова таблиці зі структурованими даними для подальшої моделі аналізу);
- побудова моделі та її оцінка (коригування параметрів побудови моделі; порівняння отриманої моделі з задачею, яку вона повинна вирішити);
- використання отриманих знань (використання отриманих результатів для подальшої роботи з ними; вивчення специфіки використаної моделі).

Існує багато методів та алгоритмів *Data Mining*. Їх можна поділити на три групи – технологічні, статистичні та кібернетичні. Розглянемо їх детальніше на рис. 3.

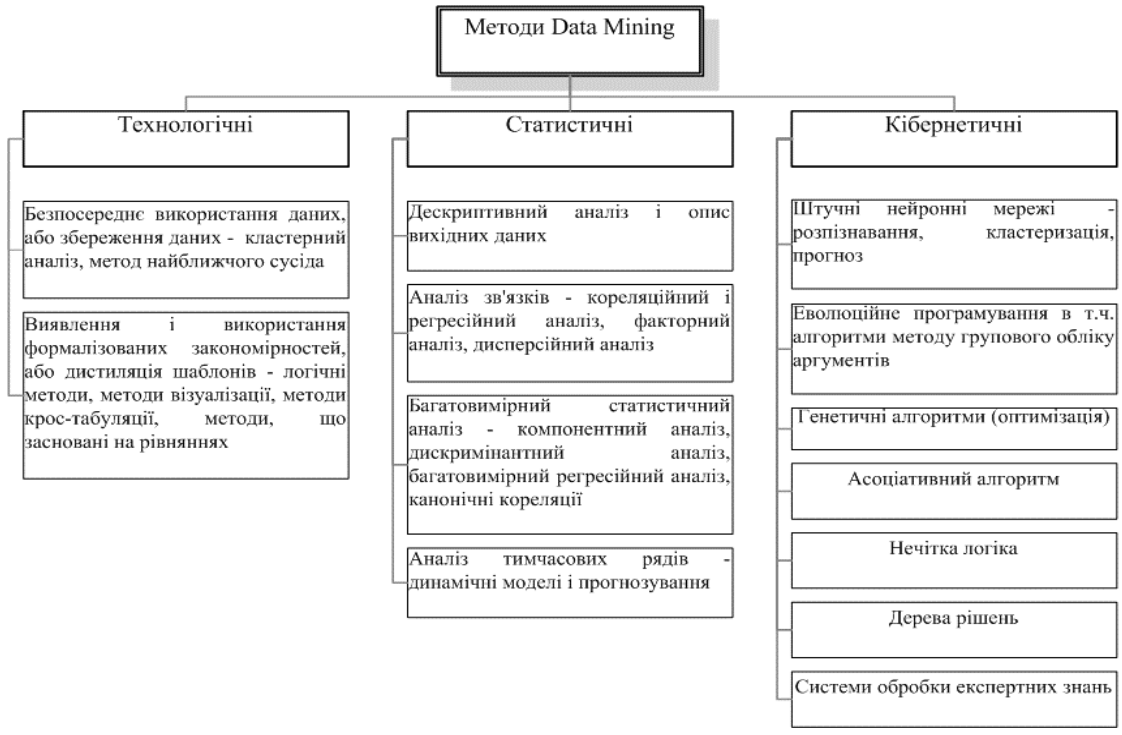


Рис. 3. Методи Data Mining

Варто зазначити, що більшість методів *Data Mining* була розроблена у межах теорії штучного інтелекту. Єдиної думки щодо того, які задачі необхідно зараховувати до *Data Mining*, немає. Більшість авторитетних джерел перераховує такі: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків. Розглянемо їх більш детально [5].

Класифікація (*Classification*). В результаті розв'язання задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних – класи; за цими ознаками новий об'єкт можна зарахувати до того чи іншого класу. Для розв'язання задачі класифікації можуть використовуватися методи: найближчого сусіда (*Nearest Neighbor*); к-ближнього сусіда (*к-Nearest Neighbor*); байєсових мереж (*Bayesian Networks*); індукції дерев рішень; нейронних мереж (*neural networks*).

Кластеризація (*Clustering*). Особливість кластеризації полягає у тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи. Прикладом методу задачі кластеризації є особливий вид нейронних мереж (карти Кохонена), що самоорганізовуються без вчителя.

Асоціація (*Associations*). У процесі розв'язання задачі пошуку асоціативних правил відшукуються закономірності між зв'язаними подіями в наборі даних. Відмінність асоціації від двох попередніх задач Data Mining: пошук закономірностей здійснюється не на основі властивостей об'єкта, що аналізується, а між кількома подіями, які відбуваються одночасно. Найвідоміший алгоритм розв'язку задачі пошуку асоціативних правил – алгоритм *Apriori*.

Послідовність (*Sequence*), або послідовна асоціація (*sequential association*). Послідовність дає змогу знайти тимчасові закономірності між транзакціями. Задача послідовності подібна до асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто, що відбуваються з деяким певним інтервалом у часі). Цю задачу *Data Mining* також називають задачею знаходження послідовних шаблонів (*sequential pattern*). Правило послідовності: після події *X* через певний час відбудеться подія *У*.

Прогнозування (*Forecasting*). В результаті розв'язання задачі прогнозування на основі особливостей існуючих даних оцінюються пропущені або ж майбутні значення цільових числових показників. Для розв'язання таких задач широко застосовуються методи математичної статистики, нейронні мережі тощо.

Візуалізація (*Visualization, Graph Mining*). В результаті візуалізації створюється графічний образ аналізованих даних. Для розв'язання задачі візуалізації використовуються графічні методи, що показують наявність закономірностей у даних. Приклад методів візуалізації – представлення даних в *2D*- і *3D*-вимірюваннях.

Підведення підсумків (*Summarization*) – задача, мета якої – опис конкретних груп об'єктів з аналізованого набору даних тощо. [6, 7]

Створення нових надтвердих матеріалів вимагає глибокого і детального аналізу доробку попередніх наукових поколінь видатних вчених. Це дуже великий масив інформації і для цього необхідно аналізувати багато наукометричних баз даних (*Web of Science, Scopus, Web of Knowledge, Google Scolar*), що викликає ряд складнощів, зокрема, значно зростає час, необхідний для обробки запитів; можуть виникати проблеми з підтримкою різних форматів даних, а також з їх кодуванням; неможливість аналізу тривалих рядів ретроспективних даних і т. ін.

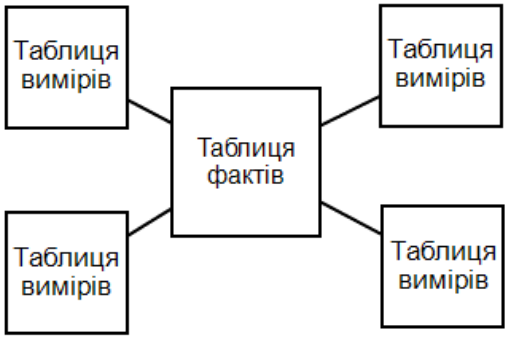
Таку проблему можна вирішити, створивши сховище даних, завданням якого буде інтеграція, актуалізація та узгодження оперативних даних з різних джерел. Саме на основі таких сховищ і здійснюються аналітична обробка і *Data Mining.*

Існує два різних підходи до проектування сховищ даних, які сформулювали двоє найбільш визначних архітекторів цих сховищ:

- Ральф Кімбалл стверджує, що сховище даних – це лише поєднання різних вітрин даних, які полегшують аналіз цих даних. Використовується підхід «знизу вгору»;

- Білл Інмон стверджує, що сховище даних є централізованим сховищем усіх даних, тобто спочатку створюється нормалізована модель сховища даних, а потім на її основі створюються вітрини даних. Це підхід «згори вниз».

Дані у сховищі можна структурувати у два способи – «зірка» та «сніжинка». Розглянемо їх більш детально.



Ри. 4. Схема типу «Зірка»

Схема «зірка» має централізоване сховище даних, яке зберігається в таблиці фактів. Схема розбиває таблицю фактів на ряд денормалізованих таблиць вимірів. Таблиця фактів містить агреговані дані, які будуть використовуватися для складання звітів, а таблиця вимірювань описує збережені дані (рис. 4).

Денормалізовані проекти менш складні, тому що дані згруповані. Таблиця фактів використовує тільки одне посилання для приєднання до кожної таблиці вимірювань. Більш проста конструкція зіркоподібної схеми значно спрощує написання складних запитів.

Схема «сніжинка» використовує нормалізовані дані – дані, усі залежності яких визначені і кожна таблиця містить мінімум надлишковості. Таким чином, окремі таблиці вимірів розгалужуються на окремі таблиці вимірів (рис. 5).

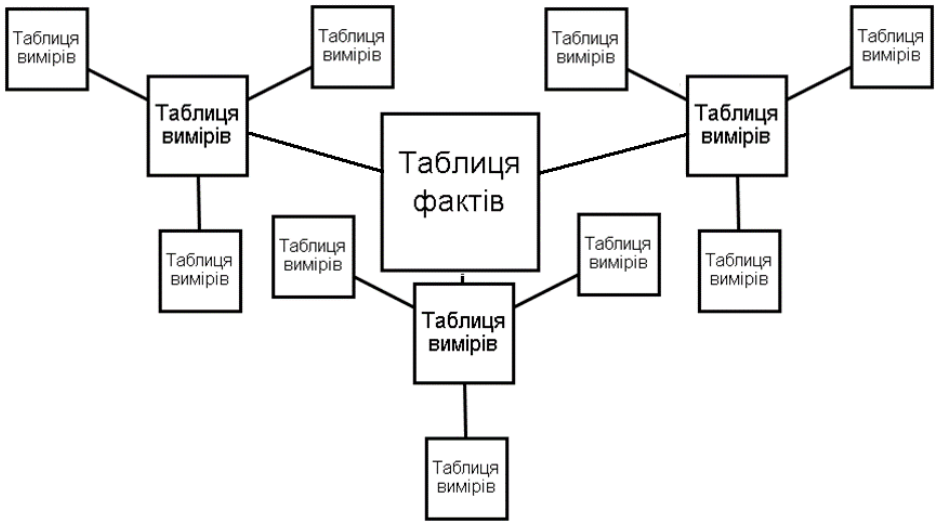


Рис. 5. Схема «сніжинка»

Схема «сніжинка» використовує менше простору на диску і краще зберігає цілісність даних. Основним недоліком є складність запитів для доступу до даних – кожен запит повинен пройти декілька з'єднань таблиць, щоб отримати відповідні дані [8].

Традиційні методи аналізу даних в основному орієнтовані на перевірку наперед сформульованих гіпотез (статистичні методи) і на «грубий розвідувальний аналіз», що становить основу оперативної аналітичної обробки даних (*Online Analytical Processing, OLAP*). Розглянемо *OLAP* більш детально.

OLAP (Online Analytical Processing) – це система аналітичної обробки даних, яка призначена для підготовки звітів, побудови прогностичних сценаріїв і виконання статистичних розрахунків на базі великих інформаційних масивів, що мають складну структуру [9]. Головна ідея даної системи полягає в побудові багатовимірних таблиць (так

званих гіперкубів), які можуть бути доступними для запитів користувачів. Ці гіперкуби будуються на основі початкових і агрегованих даних, які можуть зберігатися як в реляційних, так і в багатовимірних базах даних.

Існує три види архітектури *OLAP*-серверів:

- *MOLAP (Multidimensional OLAP)* – початкові і багатовимірні дані зберігаються в багатовимірній БД або в багатовимірному локальному кубі;
- *ROLAP (Relational OLAP)* – початкові дані зберігаються в реляційних БД або в плоских локальних таблицях на файл-сервері. Агрегатні дані можуть поміщатися в службові таблиці в тій же БД;
- *HOLAP (Hybrid OLAP)* – початкові дані залишаються в реляційній базі, а агрегати розміщуються в багатовимірній [10].

Таким чином стає зрозуміло, що актуальність *OLAP*-технологій обумовлена їх практичною значущістю для аналізу великих обсягів даних.

З іншого боку, одним з основних положень *Data Mining* є пошук неочевидних закономірностей. Інструменти *Data Mining* можуть знаходити такі закономірності самостійно і також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежності є найскладнішою задачею, перевага *Data Mining* в порівнянні з іншими методами аналізу є очевидною.

Виходячи з вищенаведеного, можна зробити висновок, що *OLAP* більше підходить для розуміння ретроспективних даних, а *Data Mining* спирається на ретроспективні дані для отримання відповідей на питання про майбутнє, тому взаємна інтеграція цих технологій повинна суттєво вдосконалити пошук і аналіз даних. Тому існує декілька варіантів такої інтеграції; розглянемо три з них у таблиці.

Варіанти інтеграції *OLAP* і *Data Mining*

Назва	Характеристика
Cubing then mining	Інтелектуальний аналіз виконується над будь-яким фрагментом будь-якої проекції гіперкуба показників або над результатами різних запитів до багатовимірних даних.
Mining then cubing	Результати інтелектуального аналізу представлені в гіперкубічній формі, яка необхідна для подальшого багатовимірного аналізу.
Cubing while mining	Автоматично виконуються однотипні механізми інтелектуальної обробки над результатом кожного кроку багатовимірного аналізу

Метою *OLAP*-аналізу є перевірка нових гіпотез, виявлення тенденцій і закономірностей, а ключова особливість *Data Mining* – нестандартність і неочевидність розшукуваних шаблонів. Засоби *Data Mining* відрізняються від *OLAP*-засобів тим, що замість перевірки заздалегідь передбачуваних гіпотез відбувається самостійне виявлення прихованих закономірностей і тенденцій, а також побудова нових гіпотез на основі знайдених взаємозв'язків. Тому інтеграція *OLAP* і *Data Mining* у систему підтримки прийняття рішень дозволить значно підвищити ефективність її використання. Ця взаємодія дає можливість аналітикам не просто відстежувати стан предметної області, а й бути в курсі неявних, неочевидних, прихованих тенденцій і закономірностей, що дозволить оцінити ефективність впровадження будь-якої технології і т. ін.

На етапі проектування аналітичної системи необхідно серйозно підходити до реалізації багатовимірного аналізу даних, тому що багато в чому це визначає ефективність, ресурсомісткість, масштабованість і інші критичні показники системи. Оптимальність використання того чи іншого способу багато в чому залежить від способу зберігання вихідних даних і визначається специфікою предметної області, для якої проектується і розробляється

аналітична система, прогнозованим обсягом інформації, що аналізується, і тими вимогами, яким повинна буде задовольняти система. Таким чином, в основі аналізу великих обсягів даних лежить багатовимірний і багатокритерійний аналіз, а підвищення ефективності аналітичних систем неможливо без інтеграції технологій *OLAP* і *Data Mining* [11].

Сферами застосування *Data Mining* є різні області знань, особливістю яких є складна системна організація. Дані в цих областях є неоднорідними, нестаціонарними, гетерогенними і великими за обсягом. Прикладами таких областей знань є медицина, молекулярна генетика, генна інженерія, прикладна хімія, астрофізика.

M.V. Dubenko, **V.M. Kulakivskyi**

V.M. Bakul Institute for Superhard Materials of National Academy of Sciences of Ukraine

BASIC PRINCIPLES OF DATA MINING

The article gives definitions of the term Data Mining, considers its methods and processes; the options for solving problems arising in the analysis of scientometric databases are given; several structures of data storage are considered; the OLAP analysis method and its capabilities are considered, a comparison is made between OLAP and Data Mining, and the areas of application of the Data Mining process are given. Analysis of this information showed that Data Mining is a necessary tool for finding new knowledge in the field of materials science and superhard materials.

Key words: *Data Mining, methods of Data Mining, stages of Data Mining, Online Analytical Processing, OLAP.*

М.В. Дубенко, **В.Н. Кулаковский**

Институт сверхтвердых материалов им. В.Н. Бакуля НАН Украины

ОСНОВНЫЕ ПРИНЦИПЫ DATA MINING

В статье даны определения термина Data Mining, рассмотрены его методы и процессы; приведены варианты решения проблем, возникающих при анализе наукометрических баз данных; рассмотрено несколько структур хранилищ данных; рассмотрен метод анализа OLAP и его возможности, проведено сравнение OLAP и Data Mining, приведены области применения процесса Data Mining. Анализ этой информации показал, что Data Mining – необходимое средство для нахождения новых знаний в сфере материаловедения и сверхтвердых материалов.

Ключевые слова: *Data Mining, интеллектуальный анализ данных, методы Data Mining, этапы Data Mining, Online Analytical Processing, OLAP.*

Література

1. Петренко А.І. Grid та інтелектуальна обробка даних Data Mining // Систем. дослідж. та інформ. технології. – 2008. – № 4. – С. 97–110.
2. Беккауер А.О. Використання технологій data mining для автоматизації бізнес-процесів на виробництві // Моделювання в економіці, організація виробництва та управління проектами. – 2016. – Вип. 1(138). – С. 161–164.
3. Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб. : BHV, 2007. – 384 с.
4. Бідюк, П.І.; Кузнєцова, Н.В. (2007). Основні етапи побудови і приклади застосування мереж Байєса. Системні дослідження та інформаційні технології. – Київ: ІПСА, 2007. – С. 26–39.
5. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – Tandem Computers Inc., 1996 – 253 p.

6. Колодчак О. М. Інтелектуальний аналіз даних // Вісник Національного університету «Львівська політехніка». Комп'ютерні системи та мережі. – 2013. – № 773. – С. 49–58.
7. Савченко Л.М., Бежитский С.С. Data Mining и области его применения // Актуальные проблемы авиации и космонавтики. – 2015. Т. 1, № 11. – С. 611–613.
8. Data Warehouse Architecture: Traditional vs. Cloud [Електронний ресурс]. – Режим доступу: https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/.
9. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.
10. Миронов А.А., Мордвинов В.А., Скуратов А.К. Семантико-энтропийное управление OLAP и модели интеграции xOLAP в SemanticNET (ONTONET). Информатизация образования и науки. – 2009. – № 2. – С. 21–29.
11. Картавая И.И. Интеграция OLAP и интеллектуальной обработки данных для анализа больших данных // Материалы X Международной студенческой научной конференции «Студенческий научный форум» [Електронний ресурс]. – Режим доступу: https://scienceforum.ru/2018/article/2018006139.

Надійшла 10.06.21

References

1. Petrenko, A.I. (2008). Grid ta intelektualna obrobka dannikh Data Mining [Grid and intelligent data processing Data Mining]. *Systemni doslidzhennia ta informatsiini tekhnolohii – Systems research and information technology, 4,* 97–110.
2. Bekkauer, A.O. (2016). Vykorystannia tekhnolohii data mining dlia avtomatyzatsii biznes-protsesiv na vyrobnytstvi [Use of data mining technologies to automate business processes in production]. *Modeliuvannia v ekonomitsi, orhanizatsiia vyrobnytstva ta upravlinnia proektamy – Modeling in economics, organization of production and project management, 1, 138,* 161–164 [in Ukrainian].
3. Barsegyan, A. A. (2007). *Tekhnolohii analiza dannykh: Data Mining, Visual Mining, Text Mining, OLAP [Data analysis technologies: Data Mining, Visual Mining, Text Mining, OLAP]*. S Pb.: BHV [in Russian].
4. Bidyuk, P.I., & Kuznetsova, N.V. (2007). *Osnovni etapy pobudovy i pryklady zastosuvannia merezh Bayesa. Sistemni doslidzhennya ta informatsiyni tehnologiyi [The main stages of construction and examples of application of Bayesian networks. Systems research and information technology]*. Kyiv: IPSA [in Ukrainian].
5. *Knowledge Discovery Through Data Mining: What Is Knowledge Discovery?* (1996). Tandem Computers Inc.
6. Kolodchak, O. M. (2013). Intelektualnyi analiz danykh [Data Mining]. Visnyk Natsionalnoho universytetu «Lvivska politekhnika». *Kompiuterni systemy ta merezhi – Bulletin of the National University «Lviv Polytechnic». Computer systems and networks, 773,* 49–58 [in Ukrainian].
7. Savchenko, L.M., & Bezhitskiy, S.S. (2015). Data Mining i oblasti eho primeneniia [DataMining and its applications]. *Aktualnye probemy aviatsii i kosmonavtiki – Actual problems of aviation and astronautics, 1, 11,* 611–613 [in Russian].
8. Data Warehouse Architecture: Traditional vs. Cloud. *panoply.io.* Retrieved from https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/.
9. Barsegyan, A. A., Kupriyanov, M. S., Stepanenko, V. V., & Holod, I. I. (2004). *Metody i modeli analiza dannykh: OLAP i Data Mining [Methods and models of data analysis: OLAP and Data Mining]*. SPb.: BHV-Peterburg [in Russian].

10. Mironov, A.A., Mordvinov, V.A., & Skuratov, A.K. (2009). Semantiko-entropiinoe upravlenie OLAP i modeli intehratsii xOLAP v SemanticNET (ONTONET) [Semantic-entropic control of OLAP and xOLAP integration models in SemanticNET (ONTONET)]. *Informatizatsiia obrazovaniia i nauki – Informatization of education and science, 2,* 21–29 [in Russian].
11. Kartavaia, I.I. (2018). Intehratsiia OLAP i intellektualnoi obrabotki dannykh dlia analiza bolshikh dannykh [Integration of OLAP and intelligent data processing for big data analysis]. Proceedings from Student Scientific Forum'18: X Mezhdunarodnaia studencheskaia nauchnaia konferentsiia – 10th International Student Scientific Conference. *scienceforum.ru*. Retrieved from https://scienceforum.ru/2018/article/2018006139 [in Russian].

УДК 004.738.5-004.732-621.391 DOI: 10.33839/2708-731X-24-1-342-346

В. М. Кулаківський, канд.техн.наук; **І. В. Скворцов**, **О. М. Давидов**, інженери

Інститут надтвердих матеріалів ім. В. М. Бакуля НАН України, вул. Автозаводська 2, 04074, м. Київ, e-mail: ivan@ism.kiev.ua

АНАЛІЗ СТАНУ ТА ПЕРСПЕКТИВИ МОДЕРНІЗАЦІЇ РОЗПОДІЛЕНОЇ ГЕТЕРОГЕННОЇ МЕРЕЖІ АКАДЕМІЧНОГО ІНСТИТУТУ НА ПРИКЛАДІ МЕРЕЖІ ІНМ НАН УКРАЇНИ

Предметом дослідження статті є поточний стан комп'ютерної мережі Інституту надтвердих матеріалів; метою дослідження є встановлення тенденцій зростання обсягів циркулюючої в ній інформації на прикладі декількох сегментів, підключених за різними технологіями, та виявлення вузьких місць існуючої топології мережі. Дано рекомендації по модернізації мережі інституту, в тому числі по використанню різних типів міжсегментних з'єднань.

Ключові слова: *ADSL, Ethernet, інтернет-трафік, швидкість з'єднання, відеоконференція.*

З кожним роком обсяг інформації, що циркулює у мережі підприємства, зростає. Це обумовлено численними причинами – багато дій переходять у он-лайн, зростає якість зображень, якість відео, змінюються звички людей, які користуються мережею.

Метою нашого дослідження є визначення загальних тенденцій зростання трафіку у мережі Інституту надтвердих матеріалів та оцінка достатності пропускної здатності деяких міжсегментних з`єднань.

Зробимо припущення, що загальні тенденції у зміні обсягу інформації у достатньо великому сегменті мережі розповсюджуються на інші сегменти. Наше припущення базується на наступних посилках: обсяг трафіку більше залежить не від користувача, а від тих джерел інформації, до яких має доступ користувач, та від їх кількості. Наприклад, за даними *HTTP Archive* [1], за період с 2010 по 2021 роки медіанне значення обсягу веб-сторінки змінилося з 467 кілобайт до 2037 кілобайт. Але ці дані не враховують збільшення популярності відеоконтенту за останні 10 років та зростання якості зображень та відеозаписів. Крім фактору збільшення обсягу веб-сторінок та збільшення мультімедіа трафіку, є і фактор, що лімітує зростання, а саме швидкість доступу до мережі. Так, більша частина мережі ІНМ побудована за технологією 100 *Mb Ethernet* (100*Base-TX*), але декілька Ethernet сегментів наукових відділів підключено за допомогою *ADSL* ліній зв'язку.