УДК 681.325

О. С. Гайденко
(аспірант кафедри «Автоматизація та комп'ютерно-інтегровані технології транспорту» Державного економіко-технологічного університету транспорту, м. Київ)

ІНТЕЛЕКТУАЛЬНА ОБРОБКА БАЗ ЗНАНЬ ГОСПОДАРСТВА ЕЛЕКТРОПОСТАЧАННЯ ЗАЛІЗНИЦЬ

Накопичена у базах знань інформація може допомогти у вирішенні завдань управління, але її ефективне використання вимагає автоматизованої обробки. У роботі проведено аналіз методів інтелектуальної обробки інформації. Виявлено, що провідним засобом обробки інформації є datamining. Запропоновано області можливого застосування алгоритмів datamining щодо баз знань господарства електропостачання залізниць.

Ключові слова: *інтелектуальна обробка інформації, datamining, knowledge discovery in databases.*

Накопленная в базах знаний информация может помочь в решении задач управления, но ее эффективное использование требует автоматизированной обработки. В работе проведен анализ методов интеллектуальной обработки информации. Выявлено, что ведущим средством обработки информации является datamining. Предложено области возможного применения алгоритмов datamining относительно баз знаний хозяйства электроснабжения железных дорог.

Ключевые слова: *интеллектуальная обработка информации, data mining, knowledge discovery in databases.*

Постановка проблеми.У сучасному світі інформаційних технологій спостерігається постійне зростання інтересу до методів інтелектуальної обробки даних. Ці тенденції визначаються, з одного боку, зростаючими обсягами інформації, що зберігається, а з іншого – постійним зростанням попиту на інформаційні послуги, пов'язані з обробкою цих даних [1].

Процеси моніторингу, автоматизованого управління та діагностики роботи систем тягового електропостачання залізниці вимагають встановлення спеціальних технічних засобів для забезпечення збору інформації з інтелектуальних датчиків та приладів комерційного обліку до баз знань. Інтелектуальна мережа електропостачання рівня публічного акціонерного товариства «Українська залізниця» передбачає об'єднання інтелектуальних мереж нижчих рівнів (залізниць, дистанцій електропостачання, тягових підстанцій) у єдиний інформаційний простір з базами знань усіх відповідних рівнів.

© Гайденко О. С., 2016

Величезний обсяг інформації, накопичений у базах знань, може приносити користь у вирішенні різноманітних управлінських завдань. Для цього необхідні ефективні засоби комплексного аналізу зібраних даних і пошуку в них закономірностей. Таке завдання через свою трудомісткість вимагає сучасних технологій обробки інформації.

Основна частина. Для вирішення деяких завдань інтелектуальної енергетики необхідна така обробка даних, яка дозволить на підставі знайдених взаємозв'язків будувати моделі, здатні описати особливості функціонування реальних систем електропостачання.

Традиційно обробка інформації з баз даних полягає в застосуванні статистичних методів для перевірки наперед сформульованих гіпотез та Online Analytical Processing (OLAP) для роботи з реляційними базами даних, а саме швидкого (оперативного) отримання відповідей на аналітичні запити. Такі підходи до аналізу даних мають низку недоліків.

Статистичні методи використовують усереднені значення. Опираючись на них, важко вияснити достеменний стан процесів у досліджуваній галузі. Усереднення по вибірці призводить до оперування фіктивними величинами і до нехтування нетипових спостережень (наприклад, пікових значень), що в свою чергу знижує достовірність результатів дослідження. Такий підхід є не ефективним для завдань, де головним критерієм кінцевих результатів є точність, наприклад, прогнозування. Зокрема, прогнозування споживання електричної енергії тяговою підстанцією критично залежить від даних, на які опирається процес, адже від точності прогнозу залежить закупівельна ціна електроенергії. Певні нетипові значення, описують унікальні, але важливі явища та можуть становити інтерес для дослідження. Навіть сама лише ідентифікація цих спостережень, може допомогти інтерпретувати сутність досліджуваних об'єктів чи явищ. Виняткові події можуть бути провідними у подальшому розвитку та поведінці складних систем як мережа тягового електропостачання.

Використання класичних методів обробки інформації, хоч і не втратило актуальності, та вузька спеціалізація не дозволяє ефективно застосовувати їх для вирішення завдань обробки баз знань господарства електропостачання залізниці.

З розвитком комп'ютерної техніки набуває популярності група технологій інтелектуального аналізу даних datamining, яка раніше не мала широкого практичного застосування в зв'язку з великою кількістю обчислень, необхідних для виконання алгоритмів. Зростання обчислювальної потужності процесорів усунуло вагомість цієї проблеми. Тепер якісний аналіз можна провести за прийнятний час [2, 3].

Накопичення великої кількості ретроспективної інформації стало поштовхом для розвитку інтелектуального аналізу. Сьогодні data mining є фактично синонімом інтелектуального аналізу даних.

Згідно визначення, data mining – це застосування певних алгоритмів для вилучення моделей із даних. Цими моделями є не відомі раніше практично корисні взаємозв'язки (тенденції, структури, залежності чи шаблони), які можуть бути інтерпретованими користувачем.

OLAP-системи забезпечують надання інформації зручної для перегляду та створення звітів, але в аналітичній обробці продукти OLAP вміють виконувати лише найпростіші дії і на практиці можуть використовуватися виключно для підготовки матеріалу [4]. Тоді як OLAP більше підходить для трактування ре-

троспективних даних, data mining використовує ретроспективні дані, щоб здобути інформацію про майбутнє. На відміну від більшості статистичних методів, data mining оперує дійсними значеннями. І в той час, як статистичні методи придатні в основному для підтвердження або спростування висунутих гіпотез, методи data mining на основі аналізу самостійно можуть будувати гіпотези, знаходити приховані правила і закономірності серед набору даних, які користувач не здатен спрогнозувати. Зважаючи на те, що найскладнішою задачею є саме формулювання гіпотези щодо взаємозв'язків, перевага data mining є очевидною.

Результати обробки даних за допомогою статистичних інструментів і OLAP-засобів можуть сприяти кращому розумінню характеру закономірностей, тому застосування data mining не виключає їх використання [3]. Інтелектуальний аналіз об'єднує останні досягнення у сфері інформаційних технологій з математичними інструментами, тому методи data mining умовно можна класифікувати як технологічні, статистичні та кібернетичні.Більшість методів datamining основані на перевірених методах з машинного навчання, розпізнавання образів і статистики. Серед багатьох методів datamining насправді існує лише кілька фундаментальних. Усі інші є доповненням або гібридами основних.

Datamining полягає у використанні технологій розпізнавання шаблонів (patterns), а також статистичних і математичних методів. Шаблони відображають фрагменти багатоаспектних взаємозв'язків у даних. Цими шаблонами є закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірки і вид розподілів значень аналізованих показників [2, 5].

Технології data miningє кроком у процесі виявлення знань у базах даних (knowledge discovery in databases (KDD)) та входять уконцепцію інформаційних сховищ даних і організації інтелектуальних обчислень.

Дані в сховищі об'єднується в цілісну структуру за різними рівнями деталізації, що надає потрібні для користувача міри узагальнення інформації. У цій концепції головне місце відведено метаданим – даним про дані. При цьому в сховищі також містяться результати перетворення інформації, її сумаризації та верифікації [6].

KDD полягає в застосуванні аналізу даних і алгоритмів виявлення, які, в прийнятних межах обчислювальної ефективності, створюють перелік шаблонів (або моделей).

Покрокове застосування KDD до баз знань господарства електропостачання залізниці умовно зображено на рис. 1. Процес KDDє інтерактивним й ітеративним, включаючи численні кроки з багатьма рішеннями, прийнятими користувачем:

1. Розвиток розуміння області застосування та визначення мети процесу KDD.

2. Створення вибірки даних: вибір набору даних, або акцентування на підмножині змінних або вибірок даних, на основі яких робитиметься відкриття.

3. Очищення та попередня обробка даних. Основні операції включають в себе видалення шуму при необхідності, збір інформації, необхідної для моделювання або обліку шуму, ухвалення рішення про стратегію обробки втраченої інформації та обліку в часовій послідовності інформації і відомих змін.

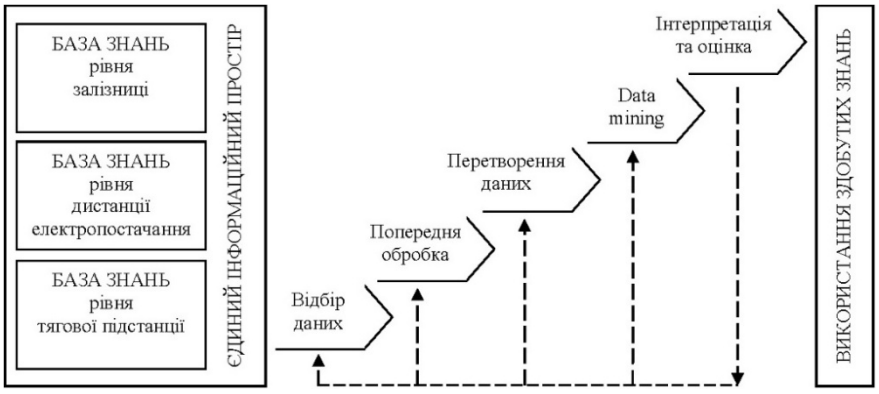

Рис. 1. **Процес KDD у базах знань інформаційного простору господарства електропостачання залізниць**

4. Зменшення даних та проектування: пошук корисних ознак для представлення даних в залежності від мети завдання. Зі зменшенням розмірності або методів трансформації, ефективна кількість змінних, відведених для компенсації, може бути зменшена, або знайдено незмінні ознаки для даних.

5. Узгодження мети процесу KDD (крок 1) з конкретним методом data mining (узагальненням, класифікацією, регресією, кластеризацією, тощо).

6. Дослідницький аналіз, моделювання і вибір гіпотези: вибір алгоритмів і методів інтелектуального аналізу даних, які використовуватимуться для пошуку шаблонів даних. Цей процес включає визначення доречних моделей та параметрів і узгодження методу data mining із загальними критеріями процесу KDD (наприклад, кінцевий користувач може бути більш зацікавлений у розумінні моделі, ніж її прогнозних можливостях).

7. Data mining: пошук моделей, які представляють інтерес, в конкретному вигляді, у вигляді дерева рішень або наборі правил класифікації, регресії, кластеризації, тощо. Користувач може суттєво допомогти методу data mining, правильно виконуючи попередні кроки.

8. Інтерпретація видобутих шаблонів, можливо, повернення до будь-якого з кроків з 1 по 7 для подальшої ітерації. Цей крок може також включати візуалізацію витягнутих шаблонів і моделей або візуалізацію даних, наданих здобутими моделями.

9. Робота з виявленими знаннями: використання знання безпосередньо, включаючи знання з іншої системи для подальших дій, або простої документації та звітності. Цей процес також включає в себе перевірку на наявність потенційних конфліктів з інформацією, яка була здобута раніше та їх вирішення.

Процес KDD може включати значні ітерації і містити петлі між будь-якими двома етапами. Основна робота KDD зосереджена на етапі 7, тим не менше, інші кроки є важливими для практичного застосування.

Після виявлення нових елементів і аспектів даних підхід до виявлення джерел і форматів даних з наступним зіставленням цієї інформації з заданим результатом може змінитися. Чим більше користувач оперуватиме даними, будуватиме моделі, оцінюватиме результати, тим кращим може бути результат. Робота з даними стає ефективнішою, коли можлива інтеграція наступних компонентів: візуалізації, графічного інструментарію, засобів формування запитів, оператив-

ної аналітичної обробки, що дозволяють зрозуміти дані й інтерпретувати результати і самі алгоритми, що будують моделі [1, 6].

Прогнозування та опис на практиці є основними завданнями datamining, мета яких може бути досягнута за допомогою ключових засобів datamining, таких як [1–3, 6–8]:

– класифікація – це віднесення об'єкту до однієї з визначених груп (класів) згідно певного набору правил або ознак, що характеризують конкретну групу. Крім того, класифікацію можна використовувати в якості вхідних даних для інших методів. Наприклад, кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

– кластеризація – загальне описове завдання, з метою визначення набору категорій або кластерів, необхідних для опису даних. Категорії можуть бути взаємно вичерпними і взаємовиключними. Досліджуючи один або більше атрибутів або класів, можна згрупувати окремі елементи даних разом, отримуючи структурований вивід. На простому рівні при кластеризації використовується один або декілька атрибутів як основа для визначення кластера схожих результатів.

– моделювання залежностей – метод знаходження моделі, яка описує істотні взаємозв'язки між змінними. Існує на двох рівнях: структурний рівень (часто у графічній формі), де змінні локально залежать одна від одної і кількісний рівень, що визначає сильні зв'язки, використовуючи деякі числові шкали.

– регресія – функція математичного очікування однієї випадкової величини залежно від значень іншої.

– підбивання підсумків– метод пошуку компактного опису для підмножини даних (наприклад, зведення до таблиці). Методи підбивання підсумків застосовується для інтерактивного дослідницького аналізу даних і автоматичної генерації звітів.

– виявлення змін і відхилень – метод, зосереджений на виявленні найістотніших змін у даних відносно норм або раніше виміряних значень.

У будь-якому алгоритмі data mining можна виділити три основних компоненти: модель представлення, модель оцінки, та пошук.

Модель представлення є мовою опису виявлених шаблонів. Якщо подання занадто обмежене, то точна модель для даних не може бути представлена, незалежно від кількості часу навчання та прикладів.

Критеріями моделі-оцінки є кількісні судження, наскільки добре певний шаблон (модель і її параметри) відповідає цілям процесу KDD. Наприклад, прогнозні моделі оцінюють тестовим набором за емпіричну точність прогнозування; описові – за розмірами прогностичної точності, новизни, корисності та зрозумілості.

Компонент пошуку в data mining складається з пошуку параметра і пошуку моделі. Після того як модель представлення і критерії моделі-оцінки нормалізовані, проблема data mining зводиться до задачі оптимізації: знайти параметри і моделі, які оптимізують критерії оцінки. Алгоритм пошуку параметра повинен шукати параметри, які оптимізують критерії моделі-оцінки, враховуючи спостережувані дані і представлення нормалізованої моделі.

Таким чином застосування алгоритмів data mining в інформаційному просторі господарства електропостачання залізниці може допомогти у виконанні завдань:

– оперативного та короткострокового прогнозування споживання електроенергії;

– виявлення невідомих взаємозв'язків між чинниками, що впливають на споживання електроенергії, та показниками споживання;

– короткострокового та довгострокового прогнозування виникнення аварійних ситуацій;

– виявлення прихованих тенденцій і закономірностей експлуатаційних процесів мережі електропостачання та прогнозування закономірностей їх розвитку;

– комплексного системного аналізу виробничих ситуацій;

– виявлення невідомих факторів впливу;

– візуалізації результатів аналізу;

– створення рекомендацій щодо прийняття управлінських рішень;

– підготовки звітів.

Висновки.

1. Визначено, що традиційні методи аналізу даних є не достатньо ефективними для їх об'єктивного застосування до баз знань інформаційного простору господарства електропостачання залізниць. Провідним засобом обробки інформації є інтелектуальний аналіз як етап процесу KDD.

2. Описано типову послідовність KDD для баз знань єдиного інформаційного простору господарства електропостачання.

3. Запропоновано області можливого застосування алгоритмів datamining.

ЛІТЕРАТУРА

1. *Джулій В. М.* Аналіз методів інтелектуальної обробки інформації / *Джулій В. М., Чешун В. М., Кривцун В. І., Солодєєва Л. В.;* ХмНУ, 2014.

2. *Петренко А. І.*Grid і інтелектуальна обробка даних DataMining // Проблемно і функціонально орієнтовані комп'ютерні системи та мережі – 2008. – №4. – С. 97-110.

3. *Текуч Н. Ю.* Актуальность и характерные особенности применения технологии datamining для решения корпоративных задач // Программные продукты и системы. – 2007. – № 4 – С. 41-43.

4. *Щавелёв Л. В.* Агрегация и интеллектуальный анализ информации хранилищ данных / *Щавелёв Л. В.,Коровкин С. Д, Левенец И. А.;* Новые информационные технологии: материалы науч.-практ. семинара / Моск. гос. ин-т электроники и математики. – М., 1998.

5. *Стеценко Д. О.* Інтелектуальна обробка даних в системі автоматизованого управління технологічним комплексом брагоректифікації / *Стеценко Д. О., Зігунов О. М., Смітюх Я. В.;*Технологический аудит. – 2014. – № 2/1(16). – С. 49-52.

6. *Шерепа Т. А.* Інформаційна технологія виділення та обробки знань у CDS/ISIS-сумісних базах даних // Бібл. вісн. – 2005. – № 5. – С. 8-13.

7. *Лєнков С. В.* Концептуальна схема системи інтелектуальної обробки даних / *Лєнков С. В.,Джулій В. М., Горбатюк О. М., Берназ Н. М.;* ХмНУ, 2014.

8. *Usama Fayyad* From Data Mining to Knowledge Discovery in Databases / *Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth;* AI Magazine – 1996. – С. 37-54.

Oles' S. Haydenko

(Graduate Student, State University for Transport Economy and Technologies)

INTELLIGENT PROCESSING KNOWLEDGE BASES OF RAILWAY POWER FACILITIES

The accumulated information in knowledge bases can help solve management problems, but its effective use requires automated processing. The analysis of methods intelligent information processing is performed in the paper. Found that the leading tool of information processing is data mining. Areas of possible data mining algorithms application to the knowledge bases of railway power facilities are suggested.

Keywords: *intelligent information processing, data mining, knowledge discovery in databases.*

REFERENCES

1. *V. Julie* Analiz metodiv intelektualnoi obrobky informazii [Analysis of intelligent information processing methods] */ V. Julie, V. Cheshun, V. Krivtsun, L. Solodeeva.;* Khmelnytsky National University*, 2014.*

2. *A. I. Petrenko* Grid i intelektualna obrobka danyh Data Mining [Grid and intellectual data processing Data Mining] // Problem and function oriented computer systems and networks – 2008. – №4. – P. 97-110.

3. *N. Y. Tekuch* Aktualnist i harakternie osobennosti primenenia tehnologii data mining dlia reshenia korporativnyh zadach [The relevance and the typical features of application technology data mining for solving corporate tasks] // Software and systems – 2007. – № 4 – P. 41-43.

4. *L. V.* Schavelev Agregacia i intelektualniy analiz informacii hranilishch danyh [Aggregation and intellectual analysis of information storage] / *L. V. Schavelev, S. D. Korovin, I. A. Levenets;* New Information Technologies: Proceedings scientific-practical workshop / Moscow State Institute of Electronics and Mathematics. – Moscow, 1998.

5. *D. O. Stetsenko* Intelektualna obrobka danih v sistemi avtomatizovanogo upravlinnya tehnologichnim bragorektifikatsiynim complexom [Intelligent data processing for automated control system of technological complex alcohol distillation and rectification] / *D. O. Stetsenko, O. M. Zigunov, A. V. Smityuh;* Technological audit – 2014. – № 2/1(16). – P. 49-52.

6. *T. A. Sherepa* Informatsiyna tehnologiya vidilennya ta obrobki znan v CDS / ISIS-sumisnih bazah danih [Information technology selection and processing of knowledge in CDS / ISIS-compliant databases] // Library Journal. – 2005. – №5. – P. 8-13.

7. *S. V. Lenkov* Conceptualna shema systemy intelektualnoi obrobki danih [Conceptual diagram of a data mining] / *S. V. Lenkov, V. N. Julie, A. M. Gorbatyuk, N. M Bernaz.;* Khmelnytsky National University, 2014.

8. *Usama Fayyad* From Data Mining to Knowledge Discovery in Databases / *Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth;* AI Magazine – 1996. – C. 37-54.