якістю. ДСТУ 2927-94. – К.: Держстандарт України,1994. – 42 с. 6. Ткаченко Н. Використання процесного підходу стандартів ISO для забезпечення регулярного виробничого циклу промислового підприємства // Вісник Нац. ун-ту «Львівська політехніка». – 2008. – №604. – С.53–57. 7. Ткаченко Н. Аспекти процесного підходу при побудові інноваційної моделі виробничого комплексу // Вісник Нац. ун-ту «Львівська політехніка». – 2008. – № 616. – С.123–128. 8. Энкарначчо Ж., Шлехтендаль Э. Автоматизированное проектирование и архитектура систем. – М.: Радио и связь, 1986. – С. 288. 9. Фомичев С., Старостина А.,Скрябина Н. Основы управления качеством. – К.: МАУП, 2002. – С. 192. 10. Сольницев Р.И., Кононюк А.Е., Кулаков Ф.М. Автоматизация проектирования ГПС. – Л: Машиностроение, Ленингр. отд-е, 1990. – С. 415. 11. Ткаченко Н.М. Критерії конкурентоспроможності продукції промислового підприємства» // Вісник Нац. ун-ту «Львівська політехніка». – 2009. – №638. – С. 72–79.

УДК 681.325

А. Батюк, С. Пилипчук, І. Цмоць
Національний університет "Львівська політехніка",
кафедра автоматизованих систем управління

ОСОБЛИВОСТІ ПОБУДОВИ ПІДСИСТЕМИ ЗБИРАННЯ, ПОПЕРЕДНЬОЇ ОБРОБКИ ТА ЗБЕРЕЖЕННЯ МЕДИЧНИХ ДАНИХ

Проаналізовано типи медичних даних, розроблено структуру підсистеми збирання, попередньої обробки та збереження даних та розглянуто основні етапи підготовки інформації до запису в сховище даних.

In this article medical data types were analyzed, gathering, preprocessing subsystems and storage was developed. Main steps of data preparation for storing in Data warehouse were described.

Постановка задачі

Основним завданням впровадження інформаційних технологій у медицину є підвищення рівня та ефективності медичної допомоги. Таке впровадження пов'язане зі створенням ієрархічних багаторівневих інформаційних систем (регіональних, територіальних, локальних і індивідуальних), які повинні забезпечувати ефективну взаємодію між рівнями систем шляхом обміну інформацією у вигляді інформаційних потоків. Для вдосконалення організаційної структури управління медичними закладами, оптимізації процесів контролю за станом здоров'я, покращання системи документообігу та автоматизації процесів одержання, збирання, збереження, пошуку та опрацювання даних необхідно впорядковувати інформаційні потоки. Впорядкування інформаційних потоків на всіх ієрархічних рівнях підвищить ефективність функціонування системи охорони здоров'я та забезпечить економічне використання кадрових, фінансових і матеріальних ресурсів [1].

У процесі лікування хворого медичні працівники створюють велику кількість різноманітної інформації (текстові записи, табличні дані, графічні зображення, цифрові сигнали), яка опрацьовується інформаційними технологіями, що об'єднуються в одну інтегровану медичну технологію:

$$IT_{MIT} = \{IT_{DW}, IT_{WEB}, IT_{ЦОС}, IT_{OLAP}, IT_{EDMS}, IT_{DM}, IT_{KDD}\},$$

де IT_{DW} – технологія інформаційних сховищ (Data Warehouse); IT_{WEB} – WEB-технології; $IT_{ЦОС}$ – технології цифрової обробки сигналів (ЦОС); IT_{OLAP} – технологія оперативної аналітичної обробки (OLAP - On-Line Analytical Processing); IT_{EDMS} – технологія автоматизації ділових процесів (EDMS – Enterprise Document Mnagement System); IT_{DM} –технологія інтелектуального аналізу даних (DM – Data Mining); IT_{KDD} – технологія, яка витягує з даних нові нетривіальні знання у формі моделей, залежностей та законів (KDD – knowledge discovery in databases). Основна роль інтегрованої медичної технології полягає в науково-практичному обґрунтуванні та знаходженні нових рішень на стику формального та логічного підходів з врахуванням емпірично описового характеру медицини. На сучасному етапі розвитку медицини мислення, логічний аналіз і використання інтегрованої медичної технології є основою клінічного діагнозу – висновку лікаря, зафіксованого на інформаційному носії, про локалізацію, характер і стадію захворювання, який обґрунтовує оптимальний вибір лікувальної тактики у межах існуючих медичних ресурсів [2].

Основною компонентою медичних інформаційних систем на всіх рівнях є підсистема збирання, попередньої обробки та збереження медичних даних. Така підсистема повинна забезпечувати: формалізацію, фільтрацію і сортування даних; автоматизоване введення документів з паперових носіїв в електронну форму; реєстрацію, облік всього обсягу вхідних, вихідних та внутрішніх документів; первинну обробку та реєстрацію документів, внесення даних у бази даних; оперативний пошук документів і пошук документів згідно з запитом за атрибутами документа (реєстраційний номер, дата, автори, виконавці тощо), ключовими словами та описами фрагментів документів; оптимальне використання та систематизацію сховищ даних відповідно до потреб інформаційних технологій; інтеграцію та взаємодію з Web-технологіями, e-mail і файловою системами; підтримку різних джерел надходження інформації; можливість роботи з сучасними СУБД [1–5].

Мета роботи полягає у розробленні структури підсистема збирання, попередньої обробки та збереження медичних даних та аналізі етапів підготовки інформації до запису в сховище даних.

Виклад основного матеріалу

Аналіз медичних даних. Значна частина медичних даних фіксується у вигляді різноманітних текстових документів (скерування на дослідження, результати аналізу, історія хвороби, рецепт, звіт про діяльність медичної установи), зображень (комп'ютерної томографії, рентгенів, ультразвукових сканерів) і цифрових сигналів (кардіографів). Приклади текстових медичних документів наведено на рис. 1, а приклад рентгенівського зображення – на рис.2.

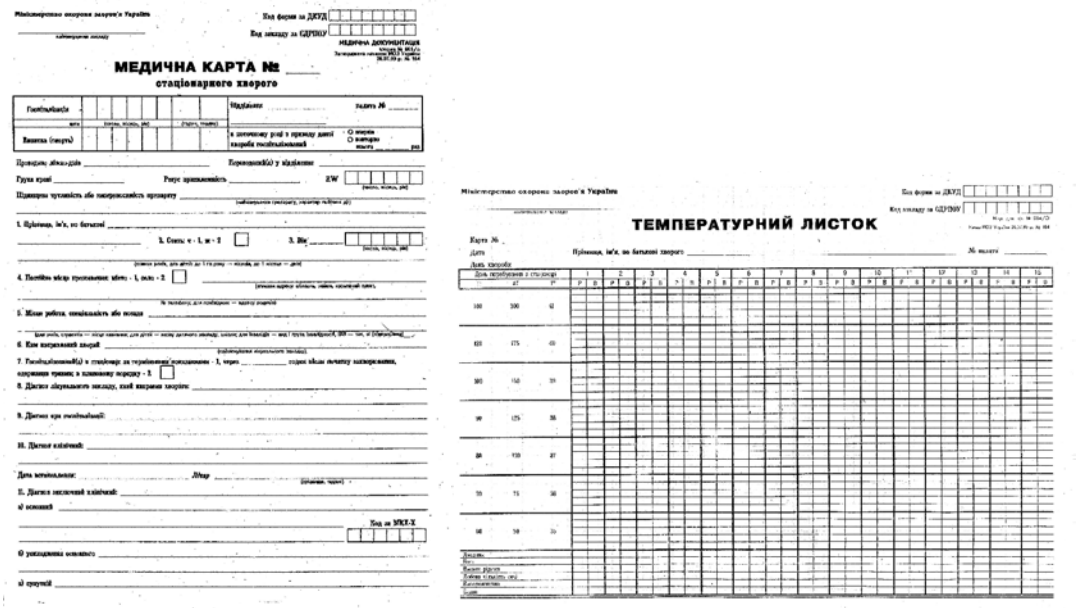

Рис. 1. Форми текстових медичних документів

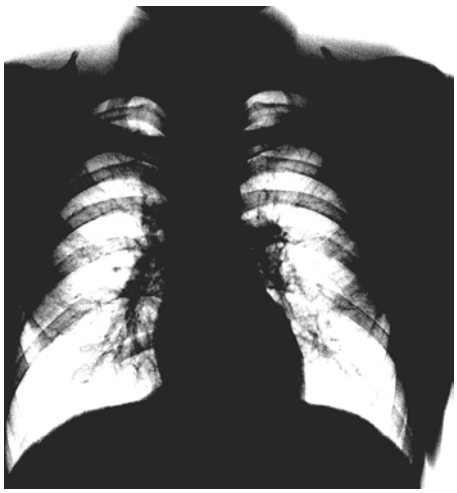

Рис. 2. Рентгенівське зображення

Звичайні медичні документи не придатні або мало придатні для безпосередньої автоматизованої обробки, оскільки вони мають складну структуру: багато розділів, пунктів, таблиць тощо. Переважно медичні документи створюються у вигляді стандартизованих історій хвороб, карт етапних епікризів, карт за окремими видами досліджень, паспортів установ охорони здоров'я. Всі ці документи мають певну форму, тобто внутрішню структуру, що відображає будову, зв'язок і спосіб взаємодії елементів об'єкта або явища, інформацію про які зафіксовано в цьому документі [1, 6].

Впродовж своєї роботи медичний фахівець заповнює відповідні стандартні форми медичних документів, в яких фіксуються такі дані:

- відомості про пацієнта: прізвище, ім'я, по батькові, рік і місце народження, характер роботи, відомості про родину;
- дані про перебіг лікування – результати лабораторних, інструментальних досліджень, призначене лікування, вплив медичних препаратів, висновки лікаря;
- дані про структуру, функцію медичних установ, лабораторні та інструментальні методи досліджень;
- статистично-управлінські дані (показники точності постанови діагнозів (відповідно до класифікації ВООЗ), тривалості перебування в стаціонарі, ступеня відновлення працездатності, розбіжності в діагнозах), які використовуються для розрахунків показників державної медичної статистики установ і показників, що характеризують роботу лікаря, відділення та установи загалом;
- економічно-планові дані про господарську і бухгалтерську діяльність медичних установ.

Медичні дані можна отримувати такими способами [6–9]:

- з аналогових носіїв шляхом розпізнавання тексту медичних документів, оцифровуванням зображень та аудіоданих;
- безпосереднім введенням даних в електронні форми документів або безпосередньо з виходів комп’ютерних томографів, рентгенів, ультразвукових сканерів і кардіографів.

Отримані одним із розглянутих способів цифрові медичні дані перед зберіганням формалізуються і фільтруються. Можливості комп’ютерного опрацювання медичних даних залежать від ступеня їх структурованості. Чим більша структурованість даних, тим більше можливостей для їх автоматизованого опрацювання. Медичні дані можна розділити на три класи:

- високоструктуровані (скерування, рецепт, результат лабораторного аналізу) мають задану структуру документа, чіткі формати і правила заповнення;
- частковоструктуровані дані (визначені деякі правила і формати, але у дуже узагальненому вигляді) – результати інструментальних досліджень та лікарські призначення;
- неструктуровані дані (інформація, записана у вигляді довільного тексту) – висновки лікарів, описи проведення процедур та інша інформація.

Структура підсистеми збирання, попередньої обробки та збереження медичних даних

Структуру підсистеми збирання, попередньої обробки та збереження медичних даних зображено на рис. 3, де ОДД – оперативне джерело даних. Основними компонентами підсистеми є засоби: класифікації даних, покращання структурованості даних, синтаксичного аналізу текстів, підготовки даних до запису у сховище даних і оперативні джерела даних. Розглянемо детальніше процеси, які передують запису інформації у сховища даних [2, 9, 10].

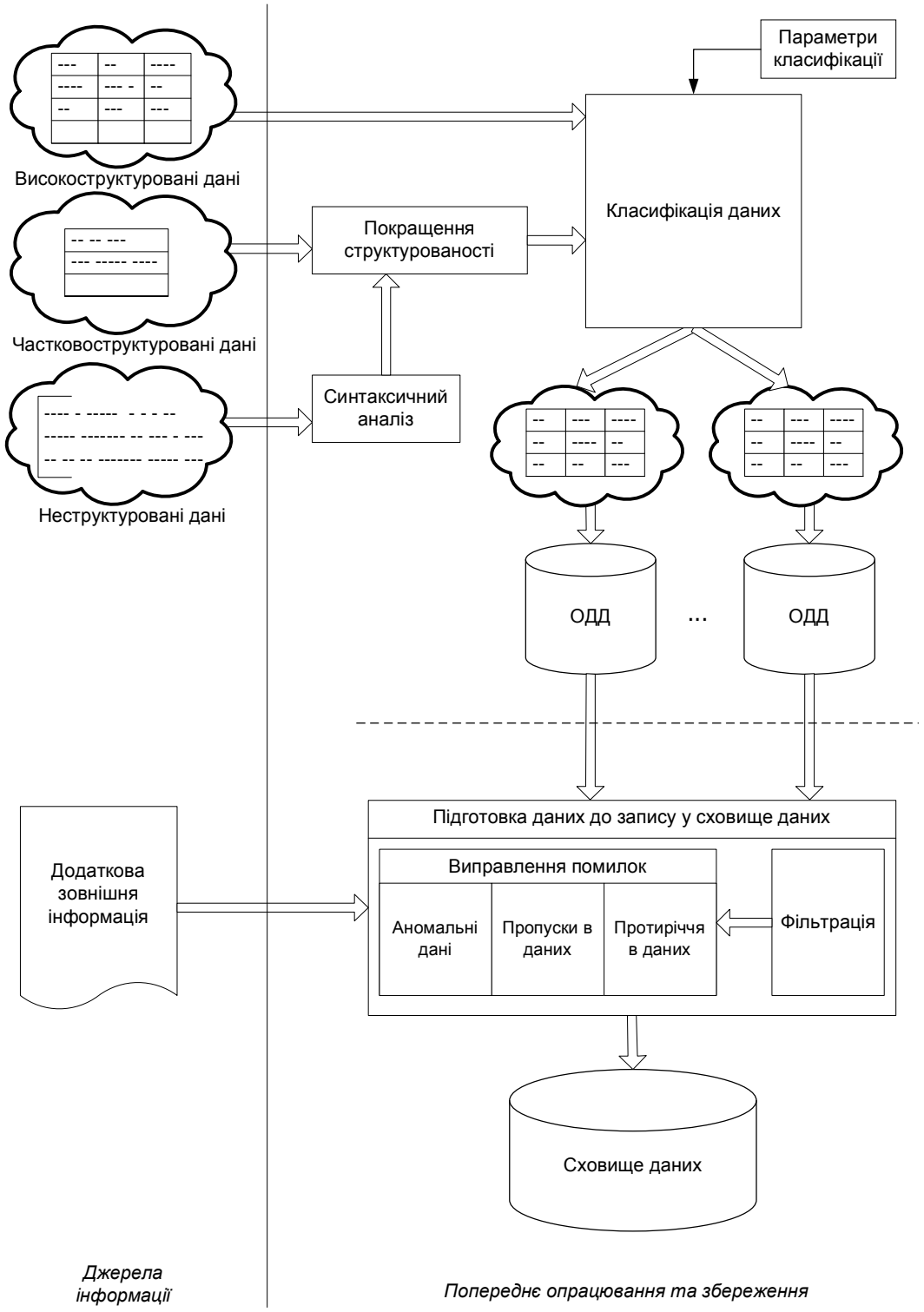

Рис. 3. Структура підсистеми збирання та збереження медичних даних

Засоби покращання структурованості даних. Однією з найважливіших умов, яка забезпечує ефективність обробки медичної інформації, є її формалізація, за якої дані, отримані з різних джерел, зводяться до однієї форми. Для виявлення певних тенденцій та закономірностей доцільно медичні дані згрупувати та подати в табличній формі. Очевидно, що механізми підготовки, фільтрації, аналізу і опрацювання високоструктурованих даних є значно продуктивнішими, ніж методи аналізу довільного тексту. Тому для ефективного використання медичних інформаційних технологій необхідно підвищувати рівень структурованості даних.

Більшість медичних даних зберігається у частковоструктурованому та неструктурованому вигляді ("Panadol gsk. Табл. 500 мг, 12т"). Такі дані необхідно очистити від можливих незначних спотворень, виділити назву препарату, виробника, лікарську форму та ін. Приклад покращання структурованості даних наведено на рис. 4.

Panadol gsk. Табл. 500 мг, 12т =>

Назва	Panadol
Виробник	GlaxoSmithKline
Форма	таблетка
Доза	500
Кількість	12

Рис. 4. Покращання структурованості даних

Засоби синтаксичного аналізу текстів застосовуються до неструктурованих текстових даних для поділу на змістовні частини і перетворення їх на документи частковоструктурованої форми.

Класифікація даних. Медична інформація може бути класифікована відповідно до дисциплінарних та проблемних властивостей, до об'єктної ознаки (лікувально-профілактична установа, матеріально-технічна база, лікувальні засоби тощо), до видів інформації (економічна, наукова, нормативно-правова тощо), до її характеру (первинна, другорядна, оперативна, оглядово-аналітична, експертна тощо). Параметри, за якими здійснюється класифікація, задаються на етапі конфігурування системи (рис. 3). Класифікувати за заданими параметрами доцільно перед збереженням даних в ОДД. Перевага класифікованих даних полягає в їх ефективнішому опрацюванні на наступних етапах роботи. Наприклад, різні типи вхідних даних можуть мати різний термін їх зберігання у ОДД, до мультимедійних даних можна застосовувати особливі алгоритми аналізу чи стиснення, деяку інформацію можна автоматично передати у зовнішні інформаційні системи (повідомлення про виявлення побічної дії препарату).

Оперативні джерела даних призначені для зберігання медичних даних у первинному, незмінному та достовірному вигляді. Збережені дані із ОДД використовуються для наповнення і зберігання даних.

Засоби підготовки даних до запису у сховище даних виконують функції фільтрації та виправлення помилок.

Фільтрування даних здійснюють кожного разу, коли необхідно відділити важливу інформацію від шуму, який її спотворює. Метою фільтрації є найточніше відтворення вихідного сигналу на фоні завади, чи визначення наявності корисного сигналу або виділення кількох сигналів, які є на вході. Відомі методи обробки сигналів [11], які широко застосовуються при обробці фізичних вимірювань (в радіолокації та радіотехніці), можна застосувати при створенні інформаційних систем аналізу даних і прогнозування процесів. Пошук взаємозалежностей і закономірностей в даних за такого підходу повинен починатись з первинної обробки даних, яка підвищить достовірність даних.

Виправлення помилок є обов'язковим етапом підготовки даних до збереження у сховищі. На цьому етапі виправляють різнотипні помилки в даних шляхом усунення суперечливості, пропусків і аномалій в даних.

Суперечливість даних усувають, видаляючи суперечливі записи або обчислюючи ймовірність кожного із суперечливих значень для збереження значення з найбільшою ймовірністю.

Пропуски в даних впливають на точність прогнозування, оскільки для забезпечення точності дані повинні надходити рівномірним постійним потоком. Заповнюють пропуски за допомогою апроксимації та визначення найімовірнішого значення. Апроксимація використовується для впорядкованих даних у випадку відсутності значення в якійсь точці. Для обчислення цього значення використовується окіл точки, результат обчислення зберігається в сховище даних. У випадку, коли неможливо визначити окіл точки, то дані для запису в сховище визначаються обчисленням найімовірнішого значення з вибірки даних.

У вибірках даних, які зберігаються в ОДД, зустрічаються аномальні значення, що сильно відрізняються від інших даних. У процесі подальшого опрацювання даних з аномаліями вони сприйматися як нормальні значення, і результат може бути спотворений. Для боротьби з цим типом помилок можна використовувати методи робастних оцінок, які характеризуються стійкістю до сильних збурень. За результатами оцінювання значення видаляється або замінюється на найближче крайнє.

Висновки

1. Основними компонентами підсистеми збирання, попередньої обробки та збереження медичних даних є засоби: покращання структурованості, синтаксичного аналізу, класифікації та підготовки інформації до запису у сховище даних.

2. Показано, що збільшення ефективності подальшого опрацювання даних досягають підвищенням їх структурованості, фільтрацією та очищенням від помилок перед збереженням у сховищі.

3. Визначено, що обов'язковим етапом підготовки даних до збереження в сховищі є виправлення помилок шляхом усунення суперечливості, пропусків і аномалій в даних.

1. Гриценко В.І., Котова А.Б., Вовк М.І., Кіфоренко С.І., Белов В.М. Інформаційні технології в біології та медицині: Курс лекцій: Навч. посібник. – К.: Наук. думка, 2007. – 382 с. 2. Цмоць І.Г, А.Є. Батюк, С.А. Батюк,С.І. Пилипчук. Вибір принципів побудови та розроблення узагальненої архітектури медичної інформаційної технології // Вісник Нац. ун-ту "Львівська політехніка", 2009. – № 650. – С.3–11. 3. Батюк А.Є., Батюк С.А., Пилипчик С.І., Цмоць І.Г. Компоненти інформаційних технологій для медицини // Тези міжнародної наукової конференції "Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту". Т.1. – Євпаторія, 2009. – С. 135–137. 4. Гланц С. Медико-биологическая статистика: Пер. с англ. – М.: Практика, 1999. – 459 с. 5. Батюк А.Є., Пасєка М.С. Концепція побудови і реалізація інформаційних систем, орієнтованих на аналіз даних // Технічні вісті. – 2000. – №1(10), 2(11). – С. 76–79. 6. Батюк А.Є., Чоп'як В.В., Цмоць І.Г., Леськів М.В., Кирик В.Ю. Інформаційна система автоматизації діяльності медичного центру клітинної імунології та алергології // Збірник наукових праць ІПМЕ НАН України: "Моделювання та інформаційні технології". – 2002. – Вип.14. – С. 175–182. 7. Чубакова И.А. Data Mining: Учеб. пособие. – БИНОМ, Лаборатория знаний, 2008. – 382 с. 8. Тарасов В.А., Герасимов Б.М., Левин И.А., Корнейчук В.А. Интеллектуальные системы поддержки принятия решений: Теория, синтез, эффективность. – К.: МАКНС, 2007. – 336 с. 9. Лук'янова В.В. Комп'ютерний аналіз даних: Посібник. – К.: Видавничий центр «Академія», 2003. – 344 с. 10. Борсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа даннях: OLAP и DataMining. – СПб.: БХВ-Петербург, 2004 – 336с. 11. Цмоць І.Г. Інформаційні технології та спеціалізовані засоби обробки сигналів і зображень у реальному часі. – Львів: УАД, 2005. – 227 с.