УДК 004.11.2

Б.А. Рудов

Академія Внутрішніх Військ МВС України, Харків

СПЕЦИФІКА ПОБУДОВИ ЕЛЕКТРОННОГО АРХІВУ ДАНИХ ВНУТРІШНІХ ВІЙСЬКОВИХ ПІДРОЗДІЛІВ

У статті проводиться аналітичний огляд сучасного електронного архіву, детально розглянута друга її складова – інформаційно-пошукова система, і її основні функціональні характеристики.

Ключові слова: *електронний архів, інформаційно-пошукова система, синхронізація даних.*

Вступ

Постановка проблеми. В умовах діяльності внутрішніх військових підрозділів відбувається накопичення значних масивів інформації, які вимагають певної синхронізації та упорядкування. Величезна кількість (до 80%) оперативної та довідкової інформації все ще залишається на паперових носіях і складується в запорошених бібліотеках або архівах.

Ручна обробка таких паперових сховищ стає вузьким ланкою функціонування багатьох установ і корпорацій.

Вирішенням проблеми може стати використання технології побудови корпоративного електронного архіву (КЕА), яка поки що достатньо нова, а її реалізація вимагає певної сміливості замовника і ставить непрості завдання перед системними інтеграторами.

Аналіз останніх досліджень і публікацій. Аналіз літературних джерел [1 – 4] по темі дослідження свідчить, що на даний момент склалася ситуація, при якій є відсутнім єдиний підхід до побудови електронного архіву даних внутрішніх військових підрозділів. Крім того, в спеціалізованій літературі відсутня концепція синхронізації електронних даних внутрішніх військових підрозділів, що викликає певні труднощі функціонування та вдосконалення електронних архівів.

Метою даної статті є аналітичний огляд методик побудови сучасного електронного архіву та обґрунтування методичних рекомендацій щодо синхронізації даних електронних архівів внутрішніх військових підрозділів.

Виклад основного матеріалу дослідження

Перш за все треба звернути увагу на те, що мова йде не про традиційну автоматизації каталогів бібліотек, а про побудову інтегрованої системи масштабу галузі або корпорації, що забезпечує ефективний доступ і зберігання величезних обсягів документів в електронному вигляді. Потреба в такій системі з'явилася досить давно і час від часу збільшувалася зрослим інтересом до відомчих і державних архівів, які містять історичну та довідкову інформації. Архіви, які працюють з паперовими бібліотечними каталогами, вже перестали забезпечувати оперативність, повноту і достовірність виконання запитів до непомірно зростаючих фондів документів. Крім цього, паперові цінності з часом приходять в непридатність і безповоротно зникають.

Величезний потік документів та інформаційних матеріалів, які обертаються всередині великих державних структур, придає новий імпульс побудові архівів електронних документів. І тут справа стосується вже не тільки компактного, безпечного зберігання і швидкого пошуку документів, але і питань оперативного аналізу, мета якого – прогнозування ринкових колізій і виявлення яких-небудь закономірностей.

Все це обумовлює актуальність створення нової інформаційної технології побудови КЕА, здатного ефективно обробляти масиви даних обсягом у сотні терабайт. Причому технологія повинна включати як засобу створення (наповнення) банку даних, так і засоби забезпечення його належного функціонування та розвитку.

КЕА можна визначити як комплекс апаратно-програмних засобів і технологій створення архіву (масштабу галузі або підприємства) документів в електронному вигляді. Мета створення КЕА полягає в забезпеченні оперативного і повноцінного доступу до всіх документів, які в ньому зберігаються і до нього надходять [2]. Для цього потрібно вирішити два основні завдання: ввести масив наявних в архіві документів і забезпечити можливість перативного повнотекстового доступу до електронних документів.

Загальну ідею можна окреслити наступним чином. Організовується розгортання високопродуктивної мережі, що включає графічні робочі станції і потужні сервери введення і обробки інформації. Для введення документів з паперових носіїв низької якості використовуються промислові сканери потокового введення і відповідні русифіковані програмні засоби. Система забезпечує ефективне індексування і повнотекстовий пошук неструктурованої інформації великого обсягу.

Основою КЕА можна вважати технології індексування та пошуку. Сьогодні намітилися кілька напрямків побудови електронних архівів в залежності від використовуваних в них методів пошуку.

Перший напрямок, іменований також корпоративним електронним архівом, відноситься до класу традиційних інформаційно-пошукових систем, заснованих на атрибутному пошуку структурованих даних. Як приклади можна привести системи побудови невеликих електронних архівів на базі програмних продуктів типу DOCS Open (PC DOCS), XDOC (Rank Xerox), SoftSolution (Novell) та ін.

Альтернативний напрям електронної архівації базується на принципі повнотекстового індексування неструктурованих даних і включає два види індексування: контекстно-незалежне індексування, яке є незалежне від природної мови з причини бінарної або словникової індексації; контекстно-залежне індексування, яке дозволяє оптимізувати індексацію та пошук з урахуванням специфіки морфології і семантики природної мови.

Відомо кілька методів контекстно-незалежного індексування. Найбільш поширений - індексація на базі інвертованої матриці, де словами або нормалізованим словоформам ставляться у відповідність адреси документів. Інший метод - бінарне індексування, наприклад, на базі теорії нейронних мереж. При використанні теорії розпізнавання образів цей метод дозволяє досягти можливості нечіткого пошуку схожих слів або "пошуку з помилками". Нечіткий пошук надає величезні можливості для виявлення слів, що містять спотворення або помилки, наприклад, текст після розпізнавання, перекладені на російську мову назви

фірм або іноземні прізвища. Однак при нечіткому пошуку користувач стикається з проблемою відсівання шуму - документів, де зустрілися подібні по синтаксису, але не за змістом слова.

В цілому технологія повнотекстового електронного архіву представлена двома основними напрямами: технологія електронного архівування, яка використовує можливості сучасних промислових СУБД, і технологія, заснована на спеціалізованих системах індексування і пошуку.

Перший підхід базується на використанні засобів відомих SQL-СУБД, типу: Oracle, Informix, Sybase та інших, здатних підтримувати сверхвеликі бази даних. Зазвичай ці СУБД мають засоби повнотекстової індексації типу інвертованої матриці. З цієї причини обсяг індексу може становити 30-350% від загального обсягу бази. Процентний розкид залежить від ступеня нормалізації індексованих слів тексту - приведення до початкової формі іменників, прикметників і дієслів. До гідності даного методу можна віднести наступне: крім функцій індексування в СУБД присутня безліч корисних і необхідних функціональних, сервісних і технологічних функцій підтримки якісної архівної діяльності і документообігу, які суттєво спрощують задачі інтегрування засобів і функцій, захисту інформації тощо; СУБД мають широке поширення, що виключає необхідність освоєння нових продуктів; дані засоби пройшли багаторічну апробацію в рамках додатків СУБД, перевірені на практиці і будуть підтримуватися і розвиватися ще досить довго.

До основних недоліків слід віднести те, що СУБД, особливо реляційного типу, споконвічно не орієнтовані на інтенсивну обробку сверхвеликого обсягу інформації. Тому ряд функцій по повнотекстовому пошуку і побудові запитів, швидкості пошуку реалізується менш ефектно і витончено, ніж в спеціалізованих пакетах.

Наприклад, більшість СУБД поки не мають засобів підтримки нечіткого пошуку. В результаті, необхідний додатковий етап верифікації введеного тексту з метою виправлення можливих помилок сканування і розпізнавання. Однак зараз виникла нова тенденція – випускаються нові модулі або версії програмних продуктів, орієнтованих на обробку надвеликих обсягів традиційних даних і даних мультимедіа.

Другий підхід, що включає повнотекстове індексування і пошук, заснований на використанні додаткових спеціалізованих пакетів повнотекстової індексації, зокрема, на базі нейронних мереж. Такі системи використовують бінарну індексацію і реалізують нечіткий пошук. Перерахуємо їх переваги: системи мають більш якісні можливості по індексуванню, пошуку і з аналізу, зокрема, вони реалізують нечіткий пошук, що дозволяє відмовитися від проблеми виправлення помилок після розпізнавання; системи характеризуються винятково високою швидкістю доступу; обсяг індексу не перевищує 30 % обсягу текстових даних; системи, в основному, підтримують різні мультимедіа дані.

Основними супутніми проблемами тут виступають наступні. По-перше, результати нечіткого пошуку прямо залежать від якості завдання запиту, і користувачі стикаються з проблемою шуму – одержання нерелевантних документів. По-друге, зазначені системи розраховані на потужні паралельні обчислювальні системи і поки не дуже поширені на платформі Intel. Але головний недолік полягає в тому, що це системи виключно індексування та пошуку – в них істотно обмежені функції управління до? Документами. На розробників покладаються дуже непрості завдання створення власних технологічних і сервісних функцій, інтегрування технологій та программно-апаратних засобів і т.п.

Незважаючи на відмінність у зазначених підходах до індексування і пошуку, можлива їх комбінація при реалізації конкретного електронного архіву.

Основними інженерно-технічними труднощами впровадження технології електронного архівування виступають вирішення двох великих задач: наповнення електронного архіву і забезпечення ефективного пошуку. Перша захоплює ряд інженерно-технічних проблем, вирішення яких може зажадати суттєвих часових витрат. Цим обумовлена важливість ефективної організації процесу розробки, що включає оптимальне планування процесів, аналіз напрацьованих технологій, створення системи управління якістю та ін Негнучкість економічної діяльності в більшості держструктур визначає обмеження на поетапність і величину зростання державних електронних архівів. Це підвищує вимоги до системного і детального проектування, створення дослідного зразка, організації випробувань і тестування. При цьому треба врахувати, що систему не можна вважати закінченою, поки не буде виконана найважливіша і найбільш трудомістка задача КЕА – введення основної частини накопиченого обсягу документів.

Цільова задача КЕА – забезпечення ефективного доступу до наявних даних з застосуванням інтелектуальних засобів. На цьому етапі найбільш актуальними є питання оптимізації запитів за критерієм швидкості виконання.

Однією з найбільш актуальних завдань організації сховищ даних є узгодження даних, узятих з різних джерел. Мова йде насамперед про спільне використання різнорідних баз даних, що не мають інтерфейсу між собою.

Розглянемо основні проблеми, що виникають

при об'єднанні множини баз даних. З технічних ческий точки зору організація великого банку даних особливих труднощів не викликає – на ринку СУБД можна вибрати підходящий продукт. Однак завдання узгодження інформаційних масивів – це вкрай дороге задоволення.

Узгодження даних потрібне в тому випадку, якщо різні бази даних мають спільні сутності. Ці сутності в різних базах даних можуть описуватися різними атрибутами і визначатися різними первинними ключами. Для узгодження даних необхідно виділити атрибути, які є загальними для одних і тих же сутностей в різних базах даних і дозволяють однозначно ідентифікувати екземпляри сутностей. Цей набір атрибутів приводиться до загального вигляду і розглядається як універсальний первинний ключ (УПК). Він дозволяє зіставляти екземпляри сутностей з різних баз даних. Основна проблема при створенні УПК міститься в тому, що узгодження даних неможливо без їх попереднього очищення. При цьому не можна забувати, що існує принципова різниця між власне очищенням даних і перевіркою їх цілісності. Під очищенням слід розуміти, по-перше, лексичну верифікацію даних, а по-друге, їх подальшу нормалізацію (тобто приведення до однакового подання).

У першому випадку мова йде, наприклад, про найпростішу перевірку типів – чи містять числові поля саме числові дані, наскільки коректно зазначені дати і т. п. Перетворення дат і часів, приведення їх до єдиного формату, з погодження атрибутів полів – все це також відноситься до лексичної верифікації. Після проведення лексичної верифікації настає черга процедури нормалізації атрибутів. Очищення даних дозволяє проводити синхронізацію сегментів інформації з різних баз даних та їх об'єднання – там, де це можливо [1].

Існуючі засоби очищення умовно можна розділити на дві категорії:

– універсальні системи, призначені для обслуговування всієї бази даних;

– верифікатори приватних атрибутів (наприклад, системи перевірки коректності імені, адреси, поштового індексу, телефону і т.п.).

На ринку представлені системи в основному категорії універсальних систем. Кожна з подібних систем по-своєму унікальна і специфічна з точки зору сфери використання. Наприклад, у продукті фірми Validy широко використовуються алгоритми нечіткої логіки, а в Apertus застосовуються правила, що записуються на спеціальній мові Object Query Language. Деякі з цих систем (наприклад, Trillium), переглядають дані в пошуках певних образів і навчаються на основі знайденої інформації. В інших образи, які підлягають розпізнаванню, задаються на етапі попереднього програмування.

Приватні верифікатори призначені для вирішення менш глобальних завдань і зазвичай обмежуються очищенням імен і адрес. Пакет компанії PostaSoft містить три бібліотеки:

– виправлення та кодування адрес;

– оформлення правильних імен;

– злиття / очищення.

На закінчення відзначимо деякі основні переваги електронного архіву. По-перше, підвищення повноти та оперативності відпрацьовування запитів до архіву. Особливо це ефективно при виконанні нестандартного нерегламентованого запиту. Відповідь, якої раніше чекали місяцями, причому без усякої впевненості, що він виявиться позитивним, тепер можна отримати за секунди й у зовсім іншій якості. По-друге, компактність і надійність зберігання. Можна відмовитися від дорогих сховищ документів, скорочувати витрати і займані площі. Звуження кола допущених осіб, контроль і облік доступу до системи дозволить підвищити не тільки збереження, але і безпека конфіденційної інформації. По-третє, створюється можливість проведення оперативного аналізу наявної інформації, що підвищить обґрунтованість рішень, приймаються вищою і середньою ланками керівників, які покладаються поки тільки на свій досвід і інтуїцію.

Висновки

Науковий результат даного дослідження представлений у формі методики синхронізації даних електронних архівів внутрішніх військових підрозділів.

Практичним результатом є запропонована структура компонент електронного архіву внутрішніх військових підрозділів.

Подальшим напрямком даного дослідження може стати проектування системи підтримки прийняття рішень стосовно побудови електронного архіву внутрішніх військових підрозділів.

Список літератури

1. Карпов В.Э. Об одной задаче очистки и синхронизации данных / В.Э. Карпов // Информационные технологии. – 2002. – № 9. – С. 33-36.

2. Клименко І.В. Система електронного документообігу в державному управлінні / І.В. Клименко, К. О. Линьов. – К. : Вид-во НАДУ, 2006. – 32 с.

3. Магомедов Г. Концепция построения электронного архива / Г. Магомедов // Информационные ресурсы России. – 2003. – № 5. – С. 16-18.

4. Матвієнко О.В. Основи організації електронного документообігу: навч. посібн. / О.В. Матвієнко, М.Н. Цивін. – К.: Центр учбової літератури, 2008. – 112 с.

Надійшла до редколегії 27.01.2012

Рецензент: д-р техн. наук, проф. Ф.В. Новіков, Харківський національний економічний університет, Харків.

СПЕЦИФИКА ПОСТРОЕНИЯ ЭЛЕКТРОННОГО АРХИВА ДАННЫХ ВНУТРЕННИХ ВОЕННЫХ ПОДРАЗДЕЛЕНИЙ

Б.А. Рудов

В статье проводится аналитический обзор современного электронного архива, детально рассмотрена вторая ее составляющая – информационно-поисковая система – и ее основные функциональные характеристики.

Ключевые слова: *электронный архив, информационно-поисковая система, синхронизация данных.*

THE SPECIFICS OF BUILDING AN ELECTRONIC ARCHIVE OF DATA INTERNAL MILITARY VODRAZDELENY

B.A. Rudov

The article presents an analytical overview of today's electronic archive, discussed in detail its second component - an information retrieval system - and its key features.

Keywords: *electronic archive of information storage and retrieval system, synchronization of data.*