УДК 004.652

А. Берко
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

МЕТОДИ ІНТЕГРАЦІЇ РЕЛЯЦІЙНИХ ТА XML-ДАНИХ У ГЕТЕРОГЕННИХ СИСТЕМАХ ЕЛЕКТРОННОГО КОНТЕНТ-БІЗНЕСУ

Розглянуто певні способи і можливості створення інтегрованих моделей даних для зберігання інформаційного наповнення в гетерогенних системах електронного контент–бізнесу, що базуються на технологіях XML. Проаналізовано проблеми структурної, синтаксичної та семантичної інтеграції даних.

Some ways and possibilities of integrated data model development for information storage of heterogeneous electronic content– business systems based on XML technologies are considered in this paper. Problems of structure, syntax and semantic transforming of data are analyzed.

Вступ. Вимоги до якості інформаційних послуг в середовищі комп'ютерних мереж невпинно зростають, а реалізація цих вимог все більше ускладнюється внаслідок неможливість простого механічного розширення обсягів інформаційного продукту, що надається користувачеві. Методологія й засоби реінжинірингу інформаційних процесів у мережах аналогічні тим, які використовуються при реінжинірингу бізнес-процесів.

Системи контент-бізнесу – це Web-системи поширення інформаційних продуктів на основі Internet-технологій. Сьогодні цей напрям є одним з найперспективніших підходів до обслуговування клієнтів як у мережі Internet, так і в інших інформаційних мережах. Особливістю таких систем є те, що продуктом їх застосування, кінцевим результатом діяльності та основним елементом функціонування є інформаційний ресурс (контент). Як правило, інформаційний ресурс систем контент-бізнесу має гетерогенний характер, що обумовлюється різнобічністю потреб і пропозицій у цій сфері діяльності. Це потребує поєднання в одному середовищі елементів баз даних, документів, web-сторінок, мультимедійних даних тощо. Отже, очевидною стає проблема інтеграції різнорідного контенту в єдиному середовищі зберігання, опрацювання та застосування. Інтеграція інформаційних ресурсів передбачає таке їх об'єднання, за якого спільне використання є простішим та ефективнішим, ніж локальне застосування кожної складової.

Рівні інтеграції даних в системах контент-бізнесу

Можна виділити три рівні інтеграції інформації в системах контент-бізнесу, кожен з яких має специфічні завдання, підходи, методи і засоби вирішення проблем.

Рівень застосування даних. На цьому рівні вирішуються завдання формування похідних, узагальнених, аналітичних даних, які отримуються шляхом застосування методів логічного виводу, аналізу, класифікації та інших методів і засобів. Такі дані утворюють узагальнені, інтегровані відомості, що дають змогу оперувати не одиничними фактами та значеннями, а цілісним описом деякої предметної галузі.

Рівень обміну даними. Проблема інтеграції даних на рівні їх переміщення полягає у формуванні однорідних інформаційних потоків на основі даних, отриманих з різноформатних джерел, при цьому мають забезпечуватись їхні цілісність та адекватність під час передавання. Сьогодні найперспективнішими засобами вирішення таких завдань є XML та JDBC. XML

забезпечує поєднання різнорідних даних у документах єдиного XML-формату. Це, своєю чергою, дає можливість переміщення та опрацювання їх в довільних середовищах, а також зворотного пересилання. JDBC – засобом обміну різнорідними даними, що ґрунтується на платформі Java. Застосування JDBC дає змогу отримувати і пересилати інтегровані дані назалежно від їх початкового формату зберігання і засобів опрацювання.

Рівень зберігання даних. Інтеграція на рівні зберігання ґрунтується, насамперед, на концепції та технологіях сховищ даних. З погляду інтеграції розрізняють такі моделі сховищ даних: федеративна, консолідована, репрезентативна, гібридна. Незалежно від моделі, процес наповнення сховища даних передбачає виконання таких кроків:

- видобування, перетворення, завантаження даних (ETL);
- агрегація та укрупнення;
- генерація метаданих;
- формування підсумкових, узагальнених, похідних та аналітичних даних.

При цьому найпринциповішим є вирішення проблем структурної, синтаксичної та семантичної інтеграції даних. Співвідношення інтеграції даних на різних рівнях показано на рис. 1.

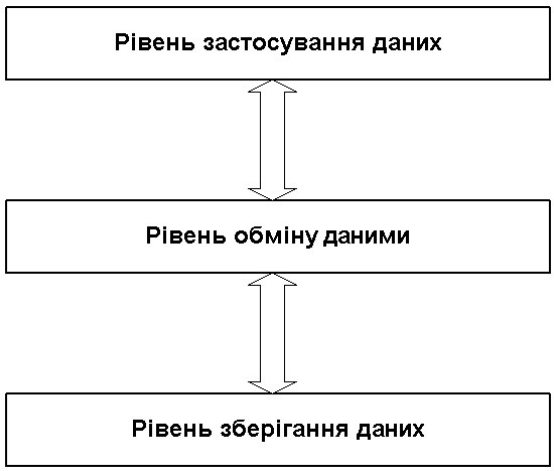

Рис 1. Рівні інтеграції даних у системах контент-бізнесу

Одним з актуальних методів вирішення задачі адекватного обміну даними на різних етапах роботи з інформаційним ресурсом системи контент-бізнесу є застосування засобів і технологій XML. Сьогодні XML визнаний як стандартизований засіб обміну даними провідними виробниками програмного забезпечення та прикладних засобів, зокрема такими, як Microsoft, IBM, Oracle, Sun та ін.

Мова XML всі частіше застосовується як формат для обміну інформацією між різними застосуваннями. Популярність XML багато в чому пояснюється його гнучкістю при поданні різних видів інформації. Із зростанням популярності XML створюється ціла серія стандартів, багато з яких були підготовані консорціумом W3C. Так, XML Schema забезпечує нотацію для визначення нових типів елементів і документів; XML Path Language (XPath) – нотацію для вибору елементів у документі XML; Extensible Stylesheet Language Transformations (XSLT) – нотацію для перетворення документів XML з одного подання в інше.

XML (*eXtensible Markup Language*) надає широкі можливості для вирішення різноманітих проблем побудови систем дистанційного навчання. Завдяки тому, що XML відокремлює структуру від відображення, стає можливим відображати той самий XML-документ як потрібно, не змінюючи при цьому сам документ. Принциповою особливістю XML є те, що предметом розмітки є не форма та синтаксис даних, а їх семантика та структура.

Основною функцією XML є визначення інших мов розмітки. XML – це метамова, а тому вона є дуже ефективним форматом подання і обміну даними. XML має багато переваг, які роблять його ефективним і зручним засобом опису даних:

1. Він зрозумілий людям у читанні й написанні, тому доступний навіть нефахівцям.

2. Це відкрита технологія. Стандарт XML запропонований W3C як платформно-незалежний засіб, на який не поширюються жодні права власності.

3. XML може застосовуватися повсюдно. Аналізатор XML можна знайти скрізь і, використовуючи відповідні інструменти, нескладно відразу ж впровадити цю технологію.

4. Мова є гнучкою. Однією з основних причин використання XML є те, що не існує чітких рамок застосування. Кожний самостійно вирішує, як використовувати його у власних застосуваннях.

5. XML є недорогим для впровадження як у великих, так і в малих організаціях.

У 2006 р. організацією W3C, що координує і виконує роботи, повязані з розвитком та запровадженням базових XML-стандартів, було затверджено вісім нових специфікацій XML, що стосуються підтримки виконання запитів, перетворення й доступу до XML-даних і документів. Найважливіші серед них: XQuery 1.0: An XML Query Language, XSL Transformations (XSLT) 2.0 і XML Path Language (XPath) 2.0.

Отже, можна зробити висновок про можливості застосування методів і технологій на основі XML як ефективний та перспективний засіб інтеграції контенту систем дистанційного навчання.

Проблема інтеграції даних при їх переміщенні

Основною метою цієї роботи є аналіз можливостей та узагальнення основних підходів перетворення даних з формату реляційних баз даних у XML-формат та зворотних перетворень. Схема переміщення контенту в системі дистанційного навчання зображена на рис. 2.

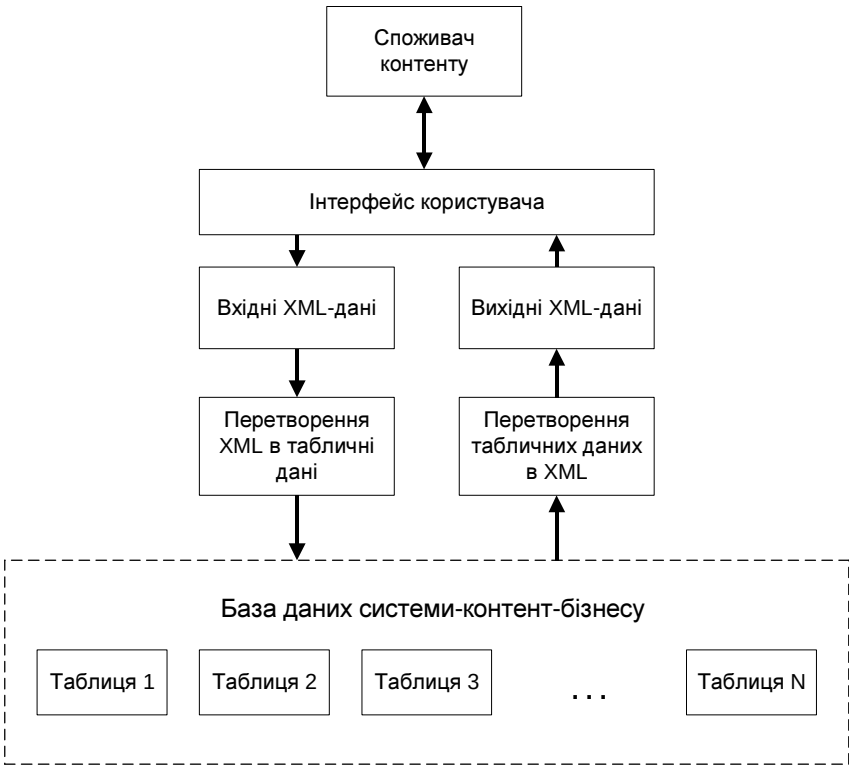

Рис. 2. Схема переміщення даних в системі контент – бізнесу

Обмін інформацією між системою та користувачем відбувається так:

- основним місцем зберігання контенту є реляційна база даних, з табличним зображенням даних;
- для передачі користувачеві табличні дані перетворюються в XML-формат;
- через інтерфейс користувача XML-документи відображаються у відповідному форматі, що відповідає завданням користувача;
- інформація, яка надходить від користувача, формується у вигляді XML- документів;
- після перетворення даних з XML-формату до табличного вони заносяться до бази даних.

Основною проблемою в діяльності такої системи є забезпечення адекватного та ефективного перетворення даних з XML-формату в реляційні дані і зворотного перетворення. Сучасні стандартні системи і технології управління базами даних не дають однозначного вирішення такої проблеми.

Способи інтеграції табличних даних та XML-документів

Серед методів і способів взаємодії баз даних з XML- даними можна виділити декілька найпоширеніших.

1. Зберігання XML- документів у спеціально визначених стовпчиках таблиць бази даних. Таким стовпчикам при визначенні надається тип XML. Цей спосіб зокрема застосовано в таких системах, як MS SQL Server 2005, Oracle, DB/2. У цьому випадку істотною перевагою є відсутність будь яких обмежень на склад і структуру XML- документів, можливість вільного обміну ними з довільними середовищами. Але при цьому таблиці бази даних з погляду структури, синтаксису та технологій стають досить неоднорідними. Кожна з таких таблиць може бути зображена як комбінація двох компонент – реляційної та XML- компоненти. Сучасні СУБД забезпечують достатні можливості щодо опрацювання XML-даних – індексування, пошук, виконання запитів, оновлення, переформатування тощо. Але при такому підході виникає проблема суміщення технологічно та методологічно різних підходів та технологій. Для опрацювання реляційної складової застосовуються SQL- орієнтовані засоби, для XML складової – XML – орієнтовані (XQuery, XPath, тощо).

Стовпчик1	Стовпчик2	...	XML- стовпчик1	XML- стовпчик2
елементарне значення1-1	елементарне значення2-1	...	XML-документ1-1	XML-документ2-1
...	...	...	...	...

Реляційна складова XML-складова

Рис. 3. Структура таблиці зі стовпчиками типу XML

Тобто процеси опрацювання таких гетерогенних таблиць стають значно складнішими за традиційні, а база даних при цьому втрачає одну з основних переваг – цілісність структури та методів роботи з даними.

2. Альтернативний підхід передбачає перетворення табличних даних до формату XML – документів для обміну та опрацювання і зворотне перетворення для зберігання. Отже, користувач фактично працює з віртуальними XML – документами, які насправді є лише зображенням даних, що зберігаються в табличній формі. В цьому випадку постає проблема впорядкування формату XML-документів, отриманих на основі табличних даних, оскільки вони не просто підлягають передачі та відображенню, але повинні забезпечити зворотне перетворення – до табличної форми. Серед способів адекватного перетворення табличних даних до формату XML можна виділити, зокрема, такі.

- Зображення таблиці як XML- документа, в якому посилання та таблицю в кореневому елементі та значення стовпчиків у рядках дочірніх елементів подаються як значення атрибутів. Механізм такого перетворення даних з табличного може бути проілюстрований таким прикладом.

Нехай задано таблицю з іменем TestTable, що визначена так:

CREATE TABLE TestTable(Col1 INT PRIMARY KEY, Col2 CHAR(15) UNIQUE, Col3 DATE)

Col1	Col2	Col3
1	x	01.01.2007
2	y	02.01.2007
3	z	03.01.2007

Рис. 4. Приклад табличного зображення даних

в XML-форматі така множина даних може бути зображена документом наступного вигляду:

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Кореневий елемент - таблиця-->
<table tblname="TestTable">
  <!-- Опис рядка-->
      <row col1="1" col2="x" col3="01.01.2007"/>
<!-- Опис рядка-->
      <row col1="2" col2="y" col3="02.01.2007"/>
<!-- Опис рядка-->
      <row col1="3" col2="z" col3="03.01.2007"/>
</table>
```

Рис. 5. Приклад зображення даних у форматіXML –документа з поданням даних як атрибутів

Отже, кожному рядку таблиці в XML–форматі відповідає елемент з порожнім значенням, атрибути якого зображають відповідні значення стовпчиків таблиці. Зворотне перетворення з XML-формату до табличного має обмежений характер, оскільки невідомими є типи і властивості окремих значень атрибутів, що відповідають значенням стовпчиків.

- Посилання на таблицю та значення стовпчиків подаються як значення елементів. Для таблиці, зображеної на рис. 4, перетворення даних до XML–формату за такою схемою дасть результат такого вигляду:

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Кореневий елемент - таблиця-->
<table> TestTable
<!-- Опис рядка-->
     <row>
          <col1>1 </col1>
          <col2>x </col2>
          <col3>01.01.2007 </col3>
     </row>
<!-- Опис рядка-->
     <row>
          <col1>2 </col1>
          <col2>y </col2>
          <col3>02.01.2007 </col3>
     </row>
<!-- Опис рядка-->
     <row>
          <col1>3 </col1>
          <col2>z </col2>
          <col3>03.01.2007 </col3>
     </row>
</table>
```

Рис. 6. Приклад зображення даних у форматіXML –документа з поданням даних як простих значень елементів

У цьому випадку структура та семантика XML- документа є ближчою до початкових даних, оскільки кожен XML-елемент утворює логічно завершену одиницю даних, аналогічну поняттю рядка таблиці. Зворотне перетворення до табличного формату, як і в попередньому випадку, носить обмежений характер. Дані з XML- документу можуть бути достатньо адекватно долучені до існуючої таблиці.

• Кореневий елемент містить посилання на таблицю та її властивості, що задаються у вигляді атрибутів, а в дочірніх елементах разом зі значеннями стовпчиків задаються їхній тип та обмеження.

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Кореневий елемент - таблиця-->
<table constraint="PRIMARY KEY(col1)"> TestTable
<!-- Опис рядка-->
    <row>
        <col1 type="INTEGER">1</col1>
        <col2 type="CHAR(15)">x</col2>
        <col3 type="DATE">01.01.2007 </col3>
    </row>
<!-- Опис рядка-->
    <row>
 <col1 type="INTEGER">2 </col1>
        <col2 type="CHAR(15)">y</col2>
        <col3 type="DATE">02.01.2007 </col3>
    </row>
<!-- Опис рядка-->
    <row>
        <col1 type="INTEGER">3</col1>
        <col2 type="CHAR(15)">z</col2>
        <col3 type="DATE">03.01.2007 </col3>
    </row>
</table>
```

Рис. 7. Приклад зображення даних у форматіXML –документу з поданням типів та значень даних

Такий формат XML-документа достатньо адекватно відображає не лише зміст та структуру початкової таблиці, а й синтаксичні особливості її елементів та додаткові властивості. Отже, на підставі цього XML-документа початкова таблиця з усіма її властивостями може бути повністю відтворена. Однак, така структура XML має значну надлишковість, оскільки опис типу повторюється багатократно для кожного з елементів, що відповідає значенню стовпчика таблиці. При значних обсягах це може створювати додаткові незручності.

• Кореневий елемент XML- документа, що зображає таблицю, містить опис її стовпчиків та обмежень , а дочірні елементи містять значення стовпчиків таблиці. При цьому для зображення службових даних застосовуються атрибути, для зображення власне даних – значення елементів.

Перетворення табличних даних, описаних вище за такою схемою, можна відобразити так.

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Кореневий елемент - таблиця-->
<table name="TestTable">
    <!-- Опис структури таблиці-->
    <tablscheme>
        <columl name="col1" type="int"/>
        <columl name="col2" type="char" length="15"/>
        <columl name="col3" type="date"/>
        <constraint name="c1" type="PRIMARY KEY(col1)"/>
        <constraint name="c2" type="unique(col2)"/>
    </tablscheme>
    <!-- Опис вмісту таблиці-->
    <tablecontent>
```

```
        <!-- Опис рядка-->
        <row>
            <rowelement name="col1">1</rowelement>
            <rowelement name="col2">x</rowelement>
            <rowelement name="col3">01.01.2007</rowelement>
        </row>
        <!-- Опис рядка-->
        <row>
            <rowelement name="col1">2</rowelement>
            <rowelement name="col2">y</rowelement>
            <rowelement name="col3">02.01.2007</rowelement>
        </row>
        <!-- Опис рядка-->
        <row>
            <rowelement name="col1">3</rowelement>
            <rowelement name="col2">z</rowelement>
            <rowelement name="col3">03.01.2007</rowelement>
        </row>
    </tablecontent>
</table>
```

Рис. 8. Приклад зображення даних у форматі XML-документа з поданням опису таблиці даних як елементів

Особливістю такого перетворення даних з табличного формату до формату XML є те, що на відміну від попередніх способів для цього існує адекватне зворотне перетворення, оскільки в тексті XML-документа є розділ, який містить опис структури таблиці, включно з іменами стовпчиків, їх типами та обмеженнями. На основі цього опису може бути згенерована таблиця бази даних, яка має відповідну структуру та властивості. У разі додавання XML-даних до наявної в базі даних таблиці, опис структури даних з XML-документа можна використати для узгодження реляційних та XML-даних.

Висновки. Як видно з викладеного вище, XML є достатньо зручним засобом у сфері розроблення систем контент-бізнесу, що забезпечує простий та ефективний обмін даними між різними компонентами системи. Тенденція орієнтувати програми на так звані "розумні" дані, що знають свій формат і легко взаємодіють з різними середовищами, відома ще з 1980-х років. Наприкінці 90-х років XX сторіччя XML розглядався як універсальний засіб для передачі актуальних транзакційних даних. Ідея полягала в тому, щоб у міру виникнення транзакцій забезпечувати постійні синхронні оновлення систем підтримки прийняття рішень. Але, хоча ця концепція й виглядає простою, у неї є ряд прихованих особливостей. По-перше, для кожної операції транзакційна система повинна генерувати документ у фіксованому форматі, а це може спричиняти додаткові часові витрати. По-друге, документи часто стають більшими за обсягом за рахунок тегів і метаданих. Наприклад, транзакції на основі протоколу XMPP [4] (extensible messaging and presence protocol - розширюваного протоколу повідомлень і присутності) містять для кожної одиниці даних теги відкриття і закриття елемента.

Незважаючи на додаткові затрати та надлишок даних, що важливо при великих обсягах транзакцій, XML залишається найдосконалішим та перспективним сьогодні засобом обміну невеликими за обсягом повідомленнями. Основним аргументом на його користь є те, що він ґрунтується на використанні текстового формату зберігання даних, що, своєю чергою, робить його незалежним від платформи. XML неподільно пов'язаний з такими технологіями, як DTD (Document Type Definition), XSL (eXtensible Stylesheet Language), XSL Transformation (мова опису перетворень документа XML), XPath (мова опису запитів до XML-Документа), XQL (аналог SQL для баз даних,

заснованих на XML) тощо. З використанням вищезазначених технологій можна спростити процеси переміщення контенту в системах, орієнтованих на поширення продуктів інформаційних технологій, сприяючи їхньому розвитку та впровадженню у різноманітних сферах.

1. Березняк Ю.Н., Скворцов В.И. Преимущества использования xml-технологий в системах дистанционного обучения через интернет.-http://nit.miem.edu.ru/cgi-bin/article?id=107. – 2005. 2. IEEE P1484.1/D6, 2000-11-1. 4. Draft Standard for Learning Technology – Learning Technology Systems Architecture (LTSA). http://edutool.com/ltsa. 3. Extensible Markup Language 1.0 (Second Edition), W3C Recommendation (6 October 2000), http://www.w3.org/TR/REC-xml. 4. A. Berko. Consolidated data models for electronic business systems. Proceedings of IXth Internationale Conference CADSM 2007. – Lviv, 2007. Pp.341–342. 5. Берко А.Ю., Висоцька В.А. Моделі інтеграції сховищ даних електронного бізнесу. АСУ та прилади автоматики // Вісник Харківського національного університету радіоелектроніки. – 2007. – № 137. – С. 127–136.

УДК 004.89

Є. Федорчук, Д. Сметана
Національний університет "Львівська політехніка",
кафедра програмного забезпечення

ПРОГРАМНІ ЗАСОБИ ДЛЯ ПОШУКУ КЛАСТЕРІВ В БАЗІ ДАНИХ ORACLE НА ОСНОВІ ТЕХНОЛОГІЇ ADO.NET

Наведено алгоритм, технологію і результати пошуку кластерів для бази даних Oracle.

Algorithm, technology and results of the searching for an clusters for Oracle database are described.

Вступ. Прикладний аналіз товарних операцій в бізнесі передбачає широке коло задач пошуку необхідного асортименту в групах товарів**,** які належать до задач кластеризації. Особливістю цих задач є їхня комбінаторна складність, обумовлена великими кількостями товарних груп та елементів у цих групах. У роботах [1, 2] подані приклади задач пошуку цінових кластерів, обгрунтовано їхнє формулювання як задач неперервної оптимізації. Подано алгоритм розв'язання означених задач на основі розбиття області обмежень. Основною проблемою для пошуку кластерів у великих базах товарних даних є формування пошукових критеріїв. В роботі розглянуто задачу проектування програмних засобів на основі технології ADO.NET для пошуку цінових кластерів в базі даних Oracle.

Аналіз задачі та алгоритму пошуку цінових кластерів. Розглянемо задачу кластеризації як задачу пошуку асортименту товарів з врахуванням їх необхідної кількості. Є групи дискретних елементів, які мають одну основну характеристику, наприклад, вартість. Необхідно вибрати таку сукупність кластер-елементів з усіх груп, щоб вона відповідала заданій вартості.

Математична постановка такої задачі як задачі неперервної оптимізації має вигляд:

$$\text{знайти} \qquad \min \Phi = \Big(\sum_{i=1}^{n} k_i c_i - C_z\Big)^2, \tag{1}$$

де n – кількість груп; k_i – кількість елементів в і-й групі; c_i – вартість і-го елемента; C_z – задана вартість набору. Обмеження задачі містять: