УДК 004.624

СТРУКТУРНО-СИНТАКСИЧНИЙ АНАЛІЗАТОРНИХ ЕЛЕКТРОНИХ ДОКУМЕНТІВ З ТАБЛИЧНИМИ СТРУКТУРАМИ

Дунько Ю.С.

Науковий керівник - доц. каф. «Системне програмне забезпечення», канд. техн. наук Блажко О.А.

Загальновідомо, що процес впровадження інформаційних систем (ІС), які керують даними в формі реляційної бази даних (БД), закінчується тільки після первинного заповнення БД на підставі вмісту множини документів організації. Трудомісткість цього процесу можна скоротити, якщо скористатися їх електронними версіями (ЕД). В роботі [1] приділяється увага автоматизації процесу перенесення вмісту ЕД в БД ІС на основі *XML*-шаблонів ЕД. Використання цієї системи переносу зменшує час оновлення БД ІС, в той же час трудомісткість створення *XML*-шаблону потребує присутності адміністратора системи, а збільшення кількості шаблонів може призводити до помилок їх вибору з боку оператора системи. Тому також необхідно зменшити трудомісткість цих процесів через їх автоматизацію. Метою цієї роботи стала автоматизація процесу визначення класу ЕД та його *XML*-шаблону. Сутність класифікації полягає у тому, що ЕД представляється як набір типів даних: слово, дата, номер, особисте ім'я та інше. ЕД має таблиці, в яких описується певний формат даних, та виділенні дані з тексту поза таблицею, які доповнюють інформацію у таблиці та мають певний формат даних. Автоматизацію пропонується проводити на основі зв'язаних процесів, схема роботи яких представлено на рисунку.

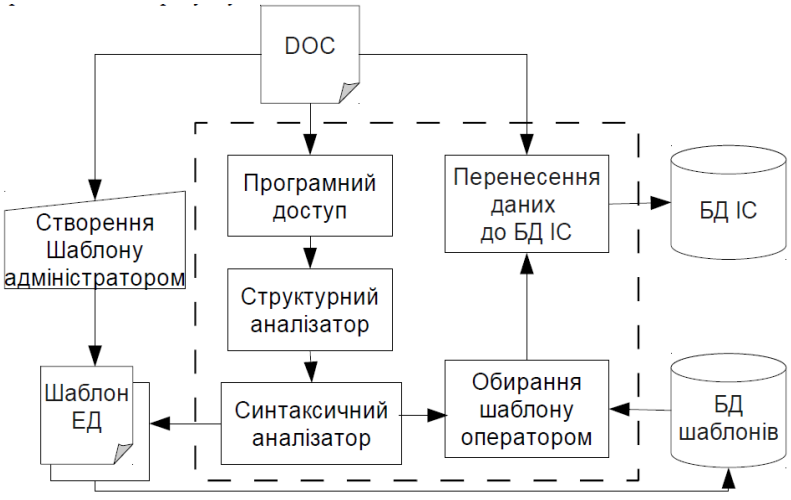

Рис. 1 Схема роботи системи автоматизації

Схема роботи модуля класифікації (МК) буде працювати в двох напрямках: при створенні шаблона та при виборі шаблона. При створенні шаблона адміністратор буде вказувати який саме ЕД описує цей шаблон, а МК буде самостійно доповнювати шаблон потрібною необхідною інформацією. А під чяс вибору шаблону МК автоматично визначає шаблони, які можуть працювати з форматами даних даного документа, зменшуючи вірогідність помилки оператора. Сьогодні існує багато синтаксичних та лексичних аналізаторів, багато з яких працюють сумісно. Найбільш розповсюдженими є *Lex*, *Yacc* та їх нащадки: *ml-Lex*, *ml-Yacc*, *Flex*, *JFlex*, *BYacc, Bison*. В роботі модуля МК синтаксичний аналізатор створено на основі *JFlex*, а лексичний аналізатор – на основі *Bison*.

Алгоритм роботи модуля МК включає наступні кроки.

Крок 1. Послідовне читання строк тексту ЕД. Крок 2. Розбивка поточної строки на прості лексеми: слово, цифра, символ пунктуації з використанням синтаксичного аналізатору *JFlex*. Крок 3. Розбір лексем та знаходження в них форматів даних з використанням лексичного аналізатору (*Bison*).

Приклад *XML*-шаблону ЕД «телефони» представлено нижче.

```
<doc>
 <head><head0>null</head0></head>
 <info><head1>"ALL_INFO$class#1$ALL_INFO"</head1></info>
 <tab_info>
  <uniteTable>"yes"</uniteTable>
  <changeHeadName>null</changeHeadName>
  <myHeadName>"position||tel1||tel2||name"</myHeadName>
  <formats><f>-</f><f>TLPHONE</f><f>TLPHONE</f><f>NAME</f></formats>
 </tab_info>
 <db_info>
   <encoding>null</encoding>
   <headType>"*:=varchar"</headType>
 </db_info>
</doc>
```

Шаблон описує документ телефонного довідника та визначає: наявність спільної інформації відносно до заголовку усього ЕД; вставляється уся інформація перед таблицею; не враховуючи першу таблицю ЕД, необхідно об’єднувати усі таблиці в ЕД в єдину, змінюючи значення їх заголовків, а дані у таблиці будуть у форматі «текст», «телефон», «телефон», «ім’я»; використовується кодова сторінка користувача та тип даних для всіх

полів «varchar». Підключення модуля МК до системи автоматизованого переносу вмісту ЕД до БД ІС дозволило спростити роботу системи, зробити обробку даних більш точною, виключаючи деякі помилки користувача, та спростити процедуру створення шаблону, що спрощує роботу адміністратора. В подальшому необхідно створити: МК із динамічно змінюючою граматикою, тобто типи даних, які може виявляти система, може розширюватися і класифікація усього документу, але для цього потрібен не тільки МК, а й текстовий аналізатор, який буде виявляти однаковість структур представлення даних для різних ЕД.

1. *Марулін С.Ю.* Автоматизація процесу обміну даними між файлами *XLS*-формату та базою даних інформаційної системи / *Блажко О.А.* / Труды десятой международной научно-практической конференции „Современные информационные и электронные технологии" – Одесса, 2009. – С. 60

2. Блажко А.А. / Алгоритм перенесения данных содержимого электронных документов *doc*-формата в базу данных информационных систем / Марулин С.Ю., Дунько Ю.С. / Труды пятой международной научно-практической конференции «Развитие научных исследований 2009» – Полтава, 2010. – С. 104–206