УДК 004.652:004.912

Я.О. Вакуленко, О.О. Мазурова

Харківський національний університет радіоелектроніки, Харків

ЗАСТОСУВАННЯ МЕТОДІВ АНАЛІЗУ ТЕКСТІВ ДЛЯ ПІДТРИМКИ КОНЦЕПТУАЛЬНОГО МОДЕЛЮВАННЯ БАЗ ДАНИХ

Робота присвячена формалізації етапу концептуального моделювання баз даних шляхом математичного описання складових моделі та їх зв'язку з сутностями та атрибутами бази даних. Математична модель доповнена статистичною мірою зустрічаємості слів – TF-індексом та враховує результати лінгвістичного аналізу вхідних документів, що описують предметну область моделювання. Запропонована модель дозволяє формалізувати підтримку на етапі концептуального моделювання з урахуванням лексем вхідного документу в якості сутностей та атрибутів бази даних. Наведено алгоритм підтримки, який може бути реалізований у складі case-засобів проектування баз даних.

Ключові слова: *база даних, концептуальне моделювання, математична модель, аналіз тексту, сутність, атрибут, частота слова.*

Вступ

Постановка проблеми. На сьогоднішній день сучасні компанії та підприємства не в змозі існувати та розвиватися без використання ефективної інформаційної системи управління та сучасних інформаційних технологій. Такі інформаційні системи (ІС) підтримають роботу в складно структурованих предметних областях, зберігають та оброблюють величезну кількість даних. Проектування баз даних (БД) є одним з найбільш відповідальних завдань, пов'язаних зі створенням таких ІС.

Аналіз предметної області та концептуальне моделювання (КМ) БД є достатньо творчим, не формалізованим та трудомістким процесом, бо включає обробку великої кількості неструктурованих текстових даних та вельми залежить від знань та досвіду проектувальника БД. Сучасні CASE-засоби не підтримують етапи аналізу та КМ, що традиційно відносяться до паперової стадії проектування БД. Таким чином, є потреба в розробці програмних продуктів, що допоможуть підтримати процес аналізу та концептуального моделювання під час розробки БД.

Аналіз основних досліджень. Концептуальне моделювання предметної області (ПО) передбачає структуроване описання різних аспектів майбутньої структури БД та вимог до функціоналу ІС [1]. КМ базується на результатах аналізу ПО, який частіше за все включає аналіз різноманітної документації в цій області. Підтримка аналізу такої текстової інформації дозволить виявити ключові слова, взаємовідносини, закономірності, які можуть бути враховані під час КМ. Це дозволить забезпечити автоматизовану підтримку КМ в складі відповідних CASE-засобів.

Для виділення ключових слів в тексті частіше використовуються статистичні [2] та синтаксичні [3] методи аналізу текстів. Статистичні методи базуються на численних даних про зустрічаємість слова у тексті. Головний їх недолік полягає в тому, що вони не враховують зв'язності тексту. Подолати цей недолік дозволяє використання синтаксичних методів. Текст підлягає обробці графематичним аналізатором [2; 4], який виконує поділ тексту на абзаци, речення і окремі слова, такі дії є необхідними для подальшої обробки. Кожне слово, виділене аналізатором, піддається морфологічному аналізу (визначається частина мови, форма, інше), виконується побудова та наповнення синтаксичних груп і виявлення відносин між ними. Таким чином, доцільно використовувати підхід, що базується на поєднанні статистичного та елементів лінгвістичного підходів [4], як найбільш ефективного для вирішення задачі аналізу тексту.

Постановка задачі. Для підтримки та часткової автоматизації процесу КМ необхідно формалізувати сам процес та розробити алгоритм його підтримки на основі статистичного та лінгвістичного аналізу документації, що описує ПО створення ІС та БД. Таким чином, була поставлена задача розробити математичну модель для формалізації процесу КМ, що враховує усі складові концептуальної моделі БД та статистичні і лінгвістичні показники аналізу вхідної документації, а також запропонувати алгоритм її використання для підтримки КМ.

Математичне описання концептуальної моделі баз даних

Процес КМ полягає в отриманні концептуальних (понятійних) складових предметної області створення БД: сутностей, їх атрибутів, зв'язків між ними, закономірностей та інших [1].

Сутність – це реальний або представляємий об'єкт ПО. Атрибут – пойменована характеристика

сутності, яка визначає його властивості. Зв'язок – асоціація, що встановлюється між сутностями і представляє собою абстракцію набору відносин між різними видами об'єктів в реальному світі.

Концептуальна модель може бути описана, як:

$$CM = \{FS, ID, SN, DM, IC, AR, LR, SR\},$$

де FS – опис функціональної структури системи, для якої розроблюється БД;

ID – опис інформаційних потреб користувачів;

SN – схема взаємозв'язку сутностей та їх атрибутів;

DM – опис документообігу системи;

IC – опис обмежень цілісності в ПО;

AR – опис алгоритмічних залежностей в ПО;

LR – опис лінгвістичних залежностей в ПО;

SR – опис вимог до ІС в цілому.

Схема взаємозв'язку SN сутностей та атрибутів

$$SN = \{E, A, RE\},$$

де $E = \{e_i(a_{ij})\}$, ($i = \overline{1,n}$, $j = \overline{1,m}$) – множина сутностей $E_i(a_{ij})$, що характеризуються атрибутами з A; $A = \{a_{ij}\}$ – множина атрибутів сутностей; $RE = \{e_k, e_l, R\}$ – множина взаємозв'язків між сутностями e_k та e_l, де R – тип зв'язку між цими сутностями.

Документообіг IC DM представляє собою множину документів, які в своїй основі містять значення атрибутів a_{ij} сутностей ПО:

$$DM = \{D_i(a_{ij})\},$$

де $D_i(a_{ij})$ – вихідний документ ІС.

Опис алгоритмічних залежностей AR в ПО передбачає моделювання залежностей $R_i(a_{ij})$ між атрибутами сутностей

$$AR = \{R_i(a_{ij})\}.$$

Інформаційні потреби користувачів $ID = \{SSF, Stat, Auto\}$ традиційно включають:

– SSF – опис потреб у пошуку $Seach(a_{ij})$, сортуванні $Sort(a_{ij})$ та фільтрації $Filter(a_{ij})$ даних за атрибутами;

– $Stat = \{St_k(a_{ij})\}$ – опис потреб в отриманні статистики на базі атрибутів сутностей;

– $Auto = \{At_k(a_{ij})\}$ – опис задач автоматизації в ПО, під час вирішення яких також враховуються ті чи інші атрибути сутностей.

Отже сутності та атрибути є найважливішими ключовими поняттями КМ. Вони присутні в описанні майже усіх складових КМ. На основі сутностей та атрибутів формуються статистичні та алгоритмічні залежності, на їх основі реалізують пошук, сортування та фільтрацію даних в БД. Тому підтримка проектувальника на етапі вибору цих ключових понять дозволить полегшити увесь процес КМ.

Доповнення математичної моделі на базі статистичних та лінгвістичного підходів до аналізу текстів

КМ базується на результатах аналізу ПО [1], що частіше складається з аналізу вхідної документації Doc. До такої корисної документації слід віднести специфікації вимог, опис бізнес-процесів. В документах можуть описуватися корпоративні або галузеві стандарти, вимоги законодавства, яким повинен задовольняти продукт. При зміні існуючої системи в старій документації можна знайти функціональність, яку треба зберегти або навпаки позбавитися.

Отже документація, що описує ПО:

$$Doc = \{Doc_k(W, Wd)\},$$

де W – набір слів, що входять до документу;

Wd – залежності або зв'язки між словами.

Для пошуку слів, які можна рекомендувати проектувальнику в якості сутностей або атрибутів БД, можна використати статистичний підхід (метод підрахування TF-індексу) [2] в поєднанні з методами синтаксичного аналізу текстів для моделювання зв'язків між сутностями та атрибутами.

Показник TF (англ. term frequency – частота слова) – статистична міра, яка використовується для оцінки важливості слова в контексті документа $Doc_k(W, Wd)$. Вона визначається як відношення числа входження n_i деякого слова W_i до загальної кількості слів $\sum_k n_k$ документа

$$tf(W_i, Doc_k) = \frac{n_i}{\sum_k n_k}.$$

Сутностями та атрибутами БД можуть виступати лексеми, що є частинами мови, а саме іменниками, та мають досить високе значення TF.

Для оцінки важливості слова також можна використовувати показник TF-IDF (від англ. TF – term frequency, IDF – inverse document frequency) – статистична міра, яка використовується для оцінки важливості слова в контексті документа, що є частиною колекції документів або корпусу.

Звісно частота зустрічаємості не бере до уваги те, що між сутністю та атрибутом повинен бути смисловий зв'язок. Такий зв'язок часто описується певними дієсловами. Наприклад, сутність може "описуватись", "характеризуватись", "мати", якийсь атрибут, а атрибут в свою чергу може "належати",

“описувати”, тощо. Чіткого списку таких дієслів не існує, але під час практичного застосування моделі множина V таких дієслів може поповнюватися.

Отже, множина сутностей E може бути обрана з множини лексем W документу $Doc_k(W, Wd)$ та розглядатися як кінцева множина слів, що належать до іменників $Wn \subset W$. Отже лексема $W_i \in W$ може бути рекомендована в якості сутності $e_j \in E$ ($E = \{e_j\}$, $j = \overline{1,m}$), в тому випадку, якщо вона є іменником $e_j \subset Wn$ та її частота зустрічаємості $tf(W_i, Doc_k) = tf(e_j) \geq tfe_{min}$, де tfe_{min} – мінімальна частота зустрічаємості сутностей у текстах.

Множина атрибутів A також може бути сформована на базі документу $Doc_k(W, Wd)$ як кінцева множина слів, що належать до іменників, а також пов’язані з сутностями зв’язками RE, що є дієсловами: $Wv \subset W$ та $Wv \subset V$. Отже лексема $W_i \in W$ може бути рекомендована в якості атрибута $a_{ij} \in A$ сутності $e_j \in E$ ($A \in W$, $A = \{a_{ij}\}$, $i = \overline{1,n}$, $j = \overline{1,m}$), в тому випадку, якщо вона пов’язана з сутністю e_j зв’язком $RE(e_j, a_{ij})$, що є певним дієсловом $RE(e_j, a_{ij}) \in V$, а також має частоту зустрічаємості $tf(W_i, Doc_k) = tf(a_{ij}) \geq tfa_{min}$, де tfa_{min} – мінімальна частота зустрічаємості атрибутів у текстах.

Для того, щоб визначити мінімальні частоти зустрічаємості сутностей tfe_{min} та атрибутів tfa_{min}, було проведено експеримент по дослідженню певної кількості документів, що описують ПО, та схем БД, що були побудовані на їх основі. За результатами експерименту для документів з кількістю слів не вище 2000 для сутностей $tfe_{min} = 5,3$, для атрибутів $tfa_{min} = 3,2$. Для документів з більшою кількістю слів значення частот tfe_{min} та tfa_{min} будуть меншими, але $tfe_{min} \succ tfa_{min}$. Ці значення можуть бути отримані шляхом накопичення статистики з аналізу документів Doc та відповідних схем БД.

Опис алгоритму підтримки КМ

На основі наведеної математичної моделі КМ, що враховує результати аналізу тексту з використанням лінгвістичного та статистичного підходів, може бути запропонований наступний загальний алгоритм підтримки процесу КМ.

1 етап: попереднє опрацьовування тексту вхідного документу $Doc_k(W, Wd)$, що містить результати аналізу ПО розробки БД, з метою виявлення множини ключових слів $Wk \subset W$, серед яких можуть бути знайдені сутності та атрибути БД. На цьому етапі проводиться графематичний аналіз [2] тексту $Doc_k(W, Wd)$ та видалення неінформативних частин, тобто стоп-слів. Кандидати в ключові слова $W_i \in Wk$ обираються серед слів, які не є стоп-словами, тобто не представляють цінності при даному типі обробки: прийменники, сполучники, вигуки, тощо.

2-й етап: розрахунок частоти зустрічаємості $tf(W_i, Doc_k)$ в документі $Doc_k(W, Wd)$ ключових слів $W_i \in Wk$. З урахуванням $tf(W_i, Doc_k)$ слово воно може бути віднесено до:

– попередньої множини сутностей $E^{/} = \{W_i\}$, якщо $tf(W_i, Doc_k) \geq tfe_{min}$;

– попередньої множини атрибутів $A^{/} = \{W_i\}$, якщо $tfe_{min} \geq tf(W_i, Doc_k) \geq tfa_{min}$.

3-й етап: підтримки проектувальника під час формування множини сутностей $E = \{e_j\}$ для БД. Під час цього проводиться синтаксичний аналіз речень [4], у складі яких є слова-кандидати на роль сутності, та визначається, до якої частини мови вони належать. Якщо слово $W_i \in E^{/}$ є іменником, воно пропонується проектувальнику в якості сутності та за його згодою може бути включено до множини сутностей БД $W_i = e_i \in E \subset E^{/}$.

4-й етап: підтримки проектувальника під час формування множини атрибутів $A = \{a_{ij}\}$ для сутностей $E = \{e_i(a_{ij})\}$ БД. Проводиться синтаксичний аналіз речень, у складі яких є слова-кандидати на роль атрибутів, та визначається, до якої частини мови вони належать. Якщо слово $W_j \in A^{/}$ є іменником, то для цього речення будується синтаксичне дерево та визначаються взаємозв’язки $RE(e_i, W_j)$ можливого атрибуту W_j з сутностями $e_i \in E$, що зустрічаються у відповідних реченнях. Якщо виявлено зв’язки $RE(e_i, W_j)$, що належать до множини дієслів V, слово W_j може бути запропоновано проектувальнику в якості атрибуту сутності, та за його згодою включено до множини атрибутів БД $W_j = a_j \in A$ для описання сутності $e_i(a_j)$.

5-й етап: підтримка проектувальника під час моделювання основних складових КМ на базі множини сутностей та атрибутів. Отже, з урахуванням виявлених сутностей, атрибутів та взаємозв’язків між ними може бути описана схема взаємозв'язку $SN = \{E, A, RE\}$, що є основою КМ. Під час моделювання таких складових КМ, як документообіг

системи $DM = \{D_i(a_{ij})\}$, алгоритмічні залежності $AR = \{R_i(a_{ij})\}$ та інформаційні потреби користувачів $ID = \{SSF, Stat, Auto\}$ проектувальнику можуть бути запропоновані сутності $e_i(a_{ij}) \in E$ та атрибути $a_{ij} \in A$, на базі яких можуть бути описані ці складові.

6-й етап: корегування показників моделі на основі виділених проектувальником сутностей та атрибутів. За результатами обраних проектувальником сутностей та атрибутів можуть бути проведенні перерахунки показників моделі:

– множина дієслів V може бути поповнена новими дієсловами, що зв'язували сутності та атрибути в реченні;

– значення мінімальних частот зустрічаємості сутностей tfe_{min} та атрибутів tfa_{min} можуть бути перераховані на основі частот зустрічаємості виділених сутностей та атрибутів.

Висновки та перспективи подальших досліджень

В роботі було запропоновано математичне описання концептуальної моделі бази даних, що може бути представлена як сукупність складових частин, що враховують такі базові поняття як сутності та атрибути. Математична модель доповнена статистичною мірою зустрічаємості слів – TF-індексом та враховує результати лінгвістичного та синтаксичного аналізу вхідних документів, що описують ПО моделювання. Запропонована модель дозволяє формалізувати підтримку на етапі концептуального моделювання. Наведений алгоритм підтримки може бути реалізований у складі case-засобів проектування баз даних, що дозволить забезпечити підтримку КМ, так званого паперового етапу проектування БД.

Список літератури

1. Диго С.М. Базы данных: проектирование и использование: ученик / С.М. Диго. – М.: Финансы и статистика, 2005. – 592 с: ил. ISBN 5-279-02571-2.

2. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. – 1988. – 24(5): 513-523.

3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.

4. Кобзарева Т.Ю. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию / Т.Ю. Кобзарева. – М.: РГГУ, 2015. – 371 с.

Надійшла до редколегії 6.01.2017

Рецензент: д-р техн. наук, проф. В.О. Філатов, Харківський національний університет радіоелектроніки, Харків.

ПРИМЕНЕНИЕ МЕТОДОВ АНАЛИЗА ТЕКСТОВ ДЛЯ ПОДДЕРЖКИ КОНЦЕПТУАЛЬНОГО МОДЕЛИРОВАНИЯ БАЗ ДАННЯХ

Я.А. Вакуленко, О.А. Мазурова

Работа посвящена формализации этапа концептуального моделирования баз данных путем математического описания составляющих модели и их связей с сущностями и атрибутами базы данных. Математическая модель дополнена статистической мерой встречаемости слов – TF-индексом и учитывает результаты лингвистического анализа входных документов, описывающих предметную область моделирования. Предложенная модель позволяет формализовать поддержку на этапе концептуального моделирования на основании учета лексем входного документа в качестве сущностей и атрибутов базы данных. Приведен алгоритм поддержки, который может быть реализован в составе case-средств проектирования баз данных.

Ключевые слова: *база данных, концептуальное моделирование, математическая модель, анализ текста, сущность, атрибут, частота слова.*

TEXT ANALYSIS METHODS APPLICATION FOR DATABASES CONCEPTUAL MODELING SUPPORT

Ya.O. Vakulenko, O.O. Mazurova

The work is dedicated to the formalization of the conceptual database modeling stage by creating the mathematical model describing the components and their relationship with entities and attributes of database. Mathematical model is complemented by statistical measure of words occurrence - TF-index and takes into account the linguistic analysis of incoming documents describing domain being modeling. The proposed model allows formalizing the user support during conceptual modeling through using the document input tokens as entities and attributes of database. The algorithm of support can be implemented as a part of case-design tools for databases.

Keywords: *database, conceptual modeling, mathematical model, text analysis, entity, attribute, frequency of words.*