Поддержка Проектирования Баз Данных

Оксана Мазурова
кафедра программной инженерии
Харьковский национальный университет
радиоэлектроники
Харьков, Украина
oksana.mazurova@nure.ua

Мария Широкопетлева
кафедра программной инженерии
Харьковский национальный университет
радиоэлектроники
Харьков, Украина
marija.shirokopetleva@nure.ua

Support for Database Design

Oksana Mazurova
Department of Software Engineering
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
oksana.mazurova@nure.ua

Mary Shirokopetleva
Department of Software Engineering
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
marija.shirokopetleva@nure.ua

***Аннотация*—Работа посвящена математическому описанию этапов проектирования базы данных, которое моделирует взаимосвязь этапов анализа предметной области, концептульного, инфологического и логического моделирования баз данных. Для моделирования связи между основными составляющими рассмотренных моделей базы данных использованы базовые понятия сущности и атрибута базы данных. На основании математического описания предложен алгоритм поддержки проектирования баз данных.**

***Abstract*—The work is devoted to the mathematical description of the database design stages, which simulates the interrelation of the stages of the domain analysis, conceptual, infologic and logical database modeling. To model the relationship between the main components of the database models examined, the basic concepts of the essence and attribute of the database are used. Based on the mathematical description, an algorithm for supporting the design of databases is proposed.**

Ключевы слова—база данных, концептуальное моделирование, ER-моделирование, логическое моделирование, сущность, атрибут

Keywords—database, conceptual modeling, ER-modeling, logical modeling, essence, attribute

I. Введение

На сегодняшний день современные компании и предприятия не в состоянии существовать и развиваться без использования эффективных информационных систем в своей деятельности. Информационные системы (ИС) поддерживают работу в сложно структурированных предметных областях, сохраняют и обрабатывают огромное количество данных. Проектирование баз данных (БД) является одной из наиболее ответственных задач, связанных с созданием таких ИС.

Анализ предметной области и концептуальное моделирование (КМ) БД является достаточно творческим, не формализованным и трудоемким процессом, так как включает обработку большого количества неструктурированных текстовых данных и весьма зависит от знаний и опыта разработчика БД.

Современные CASE-средства проектирования БД, такие как Vigual Paradigm, ERVisio, ERWin и другие, не поддерживают этапы анализа и КМ, которые традиционно относится к бумажной стадии проектирования БД. Таким образом, существует потребность в моделировании процесса проектирования БД, начиная с самых ранних этапов, и создании на основании такой модели алгоритмов и программных средств нового поколения с поддержкой всех этапов процесса проектирования БД.

II. Постановка задачи

Концептуальное моделирование БД предполагает структурированное описание различных аспектов будущей БД и разрабатываемой ИС [1, 2]. КМ базируется на результатах анализа предметной области (ПО) разработки, который чаще всего включает анализ разнообразной входной документации. Автоматизированная поддержка анализа такой документации [3] позволит выявить ключевые слова, взаимосвязи и закономерности, которые могут быть учтены во время КМ и последующих этапов проектирования.

В работе была поставлена задача промоделировать этапы проектирования БД для создания на основании такой модели алгоритма и программных средств поддержки всего процесса проектирования. Для достижения поставленной цели необходимо:

- разработать математическую модель, которая связывает основные этапы проектирования на основании базовых понятий БД;
- разработать алгоритм использования модели для поддержки процесса проектирования БД.

III. Математическое описание этапов проектирования базы данных

Проектирование БД представляет собой множество $DB = \{AN, CM, ER, LM, FM\}$, состоящее из этапов анализа ПО AN. концептуального CM, ER (инфологического) ER, логического LM и физического FM моделирования.

Анализ ПО AN традиционно проводится на основании результатов интервьюирования представителей заказчика и анализа входной документации Doc. К такой документации следует отнести спецификации требований, описания бизнес-процессов, корпоративные или отраслевые стандарты, и т.п.. Документация $Doc = \{Doc_k(W, Wd)\}$ состоит из множества таких документов, каждый из которых определяется набором слов W, входящих в документ, и зависимостями или связями между словами Wd.

Этап КМ предполагает выделение концептуальных (понятийных) составляющих ПО, в которой создается БД, а именно, сущностей, их атрибутов, связей между ними, закономерностей и т.п. [1, 2]. Все составляющие КМ могут быть промоделированы на основании базовых в области БД понятий атрибута ($A = \{a_{ij}\}$ – множество атрибутов) и сущности ($E = \{e_i(A)\}$ ($i = \overline{1,n}$) - множество сущностей $e_i(A)$ ПО).

Итак, этап концептуального моделирования БД может быть описан, как совокупность ряда концептуальных описаний:

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

где FS – функциональная структура проектируемой системы; ID – информационные потребности пользователей; SN – схема взаимосвязи объектов ПО; DM – документооборот в системе; IC – ограничения целостности ПО; AR – алгоритмические зависимости в ПО; LR – лингвистические зависимости в ПО; SR – требования к ИС в целом.

С учетом базовой роли сущностей и атрибутов в модели может быть предложено следующее описание составляющих КМ:

- функциональная структура $FS = \{Act, \Pr\}$ может быть проведена в виде use-case диаграммы языка UML, как множество акторов $Act = \{Act_i\}_{i=1,n}$ и их прецедентов $\Pr = \{\Pr_j(Act_i)\}_{j=1,m;i=1,n}$;
- информационные потребности пользователей $ID = \{ID_i\}_{i=1,n}$, где $ID_i = \{SSF_i, Stat_i, Auto_i\}$ - информационные потребности актора Act_i, которые традиционно для ИС включают: потребности $SSF_i = \{Sh_i, St_i, Fr_i\}$ в поиске $Sh_i = \{Sh_i^i(E, A)\}$, сортировке $St_i = \{St_k^i(E, A)\}$ и фильтрации $Fr_i = \{Fr_p^i(E, A)\}$ сущностей E по атрибутам A; потребности $Stat_i = \{Stat_k^i(E, A)\}$ в статистике на базе атрибутов A сущностей E; потребности $Auto_i = \{Auto_m^i(E, A)\}$ в автоматических и алгоритмических расчетах с использованием тех или иных атрибутов A сущностей E;
- схема взаимосвязи SN сущностей и атрибутов $SN = \{RE_{kl}\}$, где $RE_{kl} = \langle e_k, e_l, R_i \rangle$ – взаимосвязь между сущностями e_k и e_l, где R_i – тип связи между этими сущностями;
- документооборот системы $DM = \{Doc_i(E, A)\}$ может быть представлен как множество документов $Doc_i(E, A)$, которые содержат отчетную информацию на базе значений атрибутов A сущностей E ($DM \cap Doc$);
- алгоритмические зависимости $AR = \{AR_l(V, E, A)\}$ между некоторыми вычисляемыми атрибутами V и атрибутами A сущностей E.

В процессе концептуального проектирования БД могут быть сформулированы функциональные и ряд нефункциональных требований, которые в той или иной мере касаются объектов (сущностей) ПО:

$$SR = \{dV, dSh, dSt, dFr, dStat, dAuto, dDM, dAR, dMDB, dTR\},$$ где

- $dV = \{dV_m(E, A)\}$ – требование наличия функции отображения в ИС данных об основных сущностях E и их атрибутах A;
- dSh, dSt, dFr – требования возможности поиска, сортировки и фильтрации сущностей по их атрибутам, причем $Sh_i \subset dSh$, $St_i \subset dSt$, $FR_i \subset dFr$;
- $dStat$ – требование возможности получения статистик на базе атрибутов сущностей ($Stat_i \subset dStat$);
- $dAuto$ – требование автоматизации в ПО ($Avto_i \subset dAvto$);
- dDM – требование возможности формирования выходных отчетов ($dDM \subset DM$);

- dAR – требование формирования вычислимых полей на базе алгоритмических зависимостей ($dAR \subset AR$);
- $dMDB$ – требования к СУБД;
- - dTR – требования к текстовым редакторам для вывода отчетных документов из DM.

Сущности и атрибуты являются ключевыми понятиями КМ и служат основой для описания практически всех ее составляющих. Поэтому поддержка на этапе выбора этих ключевых понятий позволит облегчить весь процесс КМ. Для поиска слов, которые можно рекомендовать в качестве сущностей или атрибутов БД, можно использовать статистический подход (метод подсчета TF-индекса) [3, 4] в соединении с методами синтаксического анализа текстов для моделирования связи между сущностями и атрибутами. Показатель TF (англ. term frequency - частота слова) – статистическая мера, которую можно использовать для оценки важности слова в контексте документа $Doc_k(W,Wd)$. Она определяется как отношение числа вхождений n_i некоторого слова W_i к общему количеству слов $\sum_k n_k$ документа

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

Таким образом, множество сущностей E может быть выбрано из множество лексем W документа $Doc_k(W,Wd)$ с учетом их принадлежности множеству существительных Wn ($Wn \subset W$, $E \subset Wn$). Множество атрибутов A также может быть сформовано на базе документа $Doc_k(W,Wd)$ как конечное множество слов, которые принадлежат существительным, а также связаны с сущностями словами-связями из RE, которые являются глаголами Wv ($Wv \subset W$, $RE \subset Wv$).

Процесс инфологического (ER) моделирования является достаточно наглядным и полезным для следующих этапов проектирования. Сущности, атрибуты и взаимосвязи между сущностями в ER-модели могут быть описаны на основании соответствующих понятий из ранее построенной модели CM. Итак, в общем виде ER-модель БД может быть описана, как:

$$ER = \{E', A', SN'\},$$

где $E' = \{e_i'(A')\}$ ($i = \overline{1,m}$) - множество сущностей $e_i'(A')$ ($E \subset E'$); $A' = \{a_{ij}'\}$ – множество атрибутов сущностей ($A \subset A'$); схема взаимосвязи $SN' = \{RE_{kl}'\}$ сущностей, где $RE_{kl}' = \langle e_k', e_l', R_i' \rangle$ – взаимосвязь между сущностями e_k' и e_l', R_i' – тип связи между этими сущностями ($SN \subset SN'$).

В ходе логического моделирования реляционной БД отношения, их атрибуты и взаимосвязи между реляционными отношениями могут быть получены на основании соответствующих понятий из ранее построенной модели ER. При этом новые отношения, атрибуты и связи могут возникнуть в результате проведения нормализации логической модели. Логическая модель может быть описана, как множество:

$$LM = \{T, A'', SN''\},$$

где $T = \{t_i(A'')\}$ ($i = \overline{1,p}$) – множество реляционных отношений $t_i(A'')$, которые определены на атрибутах из множества A'' ($E' \subseteq T$); $A'' = \{a_{ij}''\}$ – множество атрибутов ($A' \subseteq A''$); схема взаимосвязи $SN'' = \{RT_{kl}\}$ отношений, где $RT_{kl} = \langle t_k, t_l, R_i'' \rangle$ – взаимосвязь между отношениями t_k и t_l, R_i'' – тип связи между этими отношениями ($SN' \subseteq SN''$).

При этом на основании промоделированных элементов КМ могут быть предложены рекомендации относительно нормальных форм логической модели. Так, благодаря промоделированным требованиям по фильтрации данных dFr может быть предложено выделение в справочные таблицы атрибутов, по которым предполагается фильтрация. Фактически, речь идет о преобразовании логической модели в четвертую нормальную форму. Такие справочные таблицы часто выделяют на практике для разработки эффективных программных систем с расширенными функциями фильтрации данных.

Физическая FM модель БД может быть построена с учетом особенностей СУБД, выбранной согласно требованиям $dMDB$, на основании модели LM. При этом на основании моделей предыдущих этапов проектирования БД на этапе физического моделирования может быть обеспечена поддержка формирования некоторых дополнительных составляющих физической модели БД. Так, благодаря промоделированным требованиям по поиску dSh и сортировке данных dSt может быть предложено добавление в физическую модель индексных структур на основании атрибутов, по которым предполагается поиск и сортировка.

IV. Описание алгоритма поддержки проектирования

На основании приведенного описания DB, которое учитывает основные этапы проектирования БД, может быть предложен следующий алгоритм поддержки проектированию БД.

1 этап: Поддержка этапа анализа ПО на основании входных документов $Doc_k(W,Wd)$. В ходе анализа разработчику может быть рекомендовано обратить внимание на некоторые ключевые слова $Wk \subset W$, которые потенциально могут описывать сущности ПО и их атрибуты. Кандидаты в ключевые слова $W_i \in Wk$ выбираются среди слов, которые не являются стоп-словами, т.е. не представляют ценности при данном виде обработки: предлоги, союзы и т.п. [4]. Далее проводится расчет частоты встречаемости $tf(W_i, Doc_k)$ в документе $Doc_k(W,Wd)$ ключевых слов $W_i \in Wk$. С учетом $tf(W_i, Doc_k)$ слово может быть отнесено к:

- предварительному множеству сущностей $E^* = \{W_i\}$;

- предварительному множеству атрибутов $A^* = \{W_i\}$.

2-й этап: поддержка на этапе КМ заключается в предоставлении дополнительной информации на базе результатов анализа ПО или ранее определенных составляющих КМ:

- во время описания функциональной структуры ИС FS может быть предложен предварительный список сущностей E^* для выбора множества акторов $Act = \{Act_i\}_{i=1,n}$ и построения use-case диаграммы;
- во время построения схема взаимосвязи $SN = \{E, A, RE\}$ для формирования множества сущностей $E = \{e_j\}$ могут быть предложены слова $W_i \in E^*$ из предварительного списка сущностей E^* ($W_i = e_i \in E$); для формирования множества атрибутов $A = \{a_{ij}\}$ могут быть предложены слова из предварительного множества атрибутов $W_j \in A^*$;
- во время описания документооборота системы $DM = \{Doc_i(E, A)\}$ разработчику может быть предложено множество атрибутов A сущностей E, на основании которых он может сформулировать содержимое выходного документа;
- во время описания алгоритмических зависимостей $AR = \{AR_l(V, E, A)\}$ разработчику может быть предложено множество атрибутов A сущностей E, на основании которых вычисляются вычислимые атрибуты V;
- во время описания информационных потребностей $ID = \{ID_i\}_{i=1,n}$ пользователей для каждого пользователя из множества акторов Act_i разработчику может быть предложено множество атрибутов A сущностей E, что позволит описать потребности в поиске $Sh_i = \{Sh_i^i(E, A)\}$, сортировке $St_i = \{St_k^i(E, A)\}$, фильтрации $Fr_i = \{Fr_p^i(E, A)\}$ и получении статистики $Stat_i = \{Stat_k^i(E, A)\}$; потребности в проведении автоматизации $Auto_i = \{Auto_m^i(E, A)\}$;
- на базе описанных выше составляющих КМ может быть предложена помощь в формулировании набора требований к созданию ИС $SR = \{dV, dSh, dSt, dFr, dStat, sAuto, dDM, dAR, dMDB, dTR\}$.

3-й этап: поддержка на этапе построения ER-модели $ER = \{E', A', SN'\}$ заключается в предоставлении дополнительной информации с этапа КМ, а именно: множество сущностей $E' = \{e_i'(A')\}$ может быть сформировано на базе сущностей КМ $E \subset E'$; множество атрибутов сущностей $A' = \{a_{ij}'\}$ - на базе атрибутов КМ $A \subset A'$; схема взаимосвязи $SN' = \{RE_{kl}'\}$ сущностей – на базе схемы взаимосвязи $SN \subset SN'$.

4-й этап: поддержка на этапе построения логической модели $LM = \{T, A'', SN''\}$ заключается в предоставлении дополнительной информации с этапа ER-моделирования, а именно: множество реляционных отношений $T = \{t_i(A'')\}$ может быть сформировано на базе сущностей ER-модели $E' \subseteq T$; множество атрибутов $A'' = \{a_{ij}''\}$ - на базе атрибутов ER-модели $A' \subseteq A''$; множество связей $SN'' = \{RT_{kl}\}$ между отношениями – на базе взаимосвязей из ER-модели $SN' \subseteq SN''$.

V. Выводы

В роботе было предложено математическое описание процесса проектирования БД как совокупности взаимосвязанных этапов, начиная с традиционно бумажных этапов анализа и концептуального моделирования. Основой для описания составляющих на основных этапах проектирования выступают понятия сущности и атрибута, что позволяет промоделировать некоторые достаточно творческие моменты проектирования БД. На основании предложенного математического описания разработан алгоритм, который обеспечивает информационную и методологическую поддержку разработчика в процессе проектирования БД.

Разработанные модель и алгоритм могут быть положены в основу создания нового поколения case-средств по проектированию БД с расширенным функционалом, включающим все этапы проектирования БД. Подобные средства, кроме того, имеет смысл разрабатывать под новые технологии, например, интерактивные столы, которые позволят одновременно отображать результаты работы на всех этапах проектирования БД, поддерживать поэтапную или циклическую модели жизненного цикла проектирования БД, а также обеспечивать работу коллектива разработчиков, что, в итоге, повысит эффективность процесса проектирования БД.

Литература References

[1] С. М. Диго, “Базы данных: проектирование и использование: Учебник”, Москва, Россия: Финансы и статистика, 2005.

[2] Arash Termehchy Cost-Effective Conceptual Design for Information Extraction / Arash Termehchy, Ali Vakilian, Yodsawalai Chodpathumwan, Marianne Winslett. - ACM Transactions on Database Systems, 2015, Volume 40, pp. 12:1-12:39;

[3] Т. Ю. Кобзарева, “В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию”, Москва, Россия: РГГУ, 2015.

[4] Я. О. Вакуленко, О. О Мазурова, “Застосування методів аналізу текстів для підтримки концептуального моделювання баз даних” в “Системи обробки інформації. Збірник наукових праць”, Харьков, Украина, 2017, випуск 1 (147), с. 127-13.