УДК 004.652
ББК 32.973

З.С. ЛУЧИНИН, И.Г. СИДОРКИНА

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ДОКУМЕНТО-ОРИЕНТИРОВАННОЙ БАЗЫ ДАННЫХ С ОТРАЖЕНИЕМ ОГРАНИЧЕНИЙ ЦЕЛОСТНОСТИ

Ключевые слова: *база данных, документо-ориентированная модель данных, большие данные, нереляционная модель данных, целостность данных, семантика.*

Успешная производственная деятельность предприятий, применяющих системы автоматизированного проектирования (САПР), зависит от эффективного использования информационной базы PDM систем (система управления данными об изделии), обеспечивающих управление информацией об изделии на протяжении всего цикла производства. Эффективное использование данных подразумевает, прежде всего, представление информации в форме, обеспечивающей легкость ее восприятия, однозначное ее понимание всеми участниками в течение всего жизненного цикла и простоту изменения данных по мере накопления знаний о предметной области. Эти требования распространяются на описание изделий и процессов, любую документацию, используемую в разных процедурах этапов жизненного цикла. Внедрение документо-ориентированных баз данных, получивших широкое распространение в высоконагруженных проектах, в современные САПР позволяет представлять в более естественном виде слабоформализуемые данные и уменьшить время принятия решений. В связи со слабым отражением ограничений предметной области на уровне документо-ориентированной базы данных обоснована и предложена расширенная документо-ориентированная модель данных, отражающая семантику предметной области.

Z. LUCHININ, I. SIDORKINA

MATHEMATIC MODEL OF DOCUMENT-ORIENTED DATABASE WITH REFLECTION OF INTEGRITY CONSTRAINTS

Key words: *database, data model, document-oriented data model, big data, non-relational data model, NoSQL, data integrity, semantics, computer-aided design.*

Successful production activities of enterprises that use computer-aided design (CAD) depend on the effective use of PDM systems infobase providing product information management throughout the production cycle. The effective use of data implies, first of all, the presentation of information in the form which provides its easy perception, its unambiguity for all the participants throughout the life cycle and the possibility to easily change any data with the accumulation of knowledge about the subject area. These requirements apply to the description of any products, processes, and documentation used in different stage life cycle procedures. The implementation of a document-oriented database, which is widely used in heavy loaded projects, in modern CAD system allows to represent poorly formalized data in a more natural form and to reduce the time needed for decision making. Due to weak reflection of application domain constraints on the level of document-oriented database, we propose a well-grounded enhanced document-oriented data model that reflects the semantics of the application domain.

При работе с современными многомодульными САПР приборостроения возникают естественные вопросы, связанные с построением общих баз данных, организацией структуры хранения и обработкой данных. От эффективного решения вопросов управления информацией зависит успешная производственная деятельность таких предприятий и организаций. Эффективность управления данными подразумевает представление информации в форме, обеспечивающей легкость ее восприятия и однозначное ее понимание всеми участниками жизненного цикла изделий [1]. Это требование распространяется на любое описание изделия, процесс или документацию, используемую в разных процедурах этапов

жизненного цикла производства. По мере роста информации усугубляется проблема на предприятиях и в организациях, производящих сложные изделия, в частности, с механическими и радиоэлектронными подсистемами.

Для хранения и обработки информации сложной структуры получили широкое распространение документо-ориентированные базы данных (ДОБД), реализующие нереляционный подход хранения данных. ДОБД представляют собой распределённую базу данных, нацеленную на обработку больших объемов слабоформализуемой информации. Внедрение ДОБД позволит сохранять индивидуальную информацию, связанную с конкретным изделием, тем самым поддерживать актуальность данных на высоком уровне, а также сократить временные затраты на принятие решений за счет меньшего времени обработки запросов в высоконагруженных PDM системах.

Модель данных, лежащая в основе ДОБД, слабо отражает семантику предметной области, что является существенным недостатком для подобных систем. Для решения данной проблемы предложено расширение семантической модели, которая будет отражать не только статические отношения между объектами, но и их поведение и ограничения. Таким образом, актуальными являются исследования в области представления семантических данных и поддержки целостности в системах управления данными [4], направленные в первую очередь на модель базы данных. Внедрение результатов семантического расширения модели данных позволит на ранних этапах проектирования структуры базы данных более строго описывать предметную область в терминах БД и уменьшить вероятность допустить ошибку на следующих шагах проектирования и эксплуатации БД.

Представление предметной области САПР в документо-ориентированной модели данных. Первым этапом проектирования базы данных является логическое проектирование. При проектировании модели предметная область (ПрО) представляется с использованием таких понятий, как объект, связь, свойство объекта, свойство связи.

В рамках документо-ориентированной базы данных объектом является документ, не имеющий строго описанную структуру данных. Каждый документ является моделью некоторого понятия или конкретного экземпляра сущности ПрО и предназначен для описания оригинала, его поведения и строения. Документы в зависимости от типа сущности ПрО разбиваются по коллекциям, имеющим свою информационную структуру. Экземпляром документа данного типа называется некоторое подмножество множества данных в описываемой ПрО, выделенное по имени и значению объектного идентификатора данного типа документа.

Каждый документ можно разбить на части. Каждая часть, присутствующая в документе, является носителем укрупненных свойств понятия, т.е. группирует какие-то свойства документа по семантике, назначению, принадлежности и т.д. Понятие части документа связано с возможностью управления формированием структуры экземпляра документа в момент его создания или редактирования, при этом понятие части имеет формальное толкование, соответствующее математическому определению понятия отношения. Содержимое части и ее структура могут изменяться со временем, что будет соответствовать определенному событию (изменение или появление новой ин-

формации о предметной области) в отражаемой предметной области. Этим понятие части отличается от понятия отношения, используемого в реляционных моделях баз данных. Поэтому введены понятия статическая и динамическая модель документа.

Статическая модель документа – множество упорядоченных частей, которое присутствует в любом документе от момента его порождения и до конца времени существования в системе.

Динамическая модель документа – множество упорядоченных частей, присутствующих в конкретном документе в данный момент времени. Динамическая модель может быть уникальной для каждого экземпляра любого типа документа, а также может отсутствовать вовсе.

На рис. 1 представлена концептуальная модель предметной области «изделие – компонент», выраженная в виде документов. Два документа, относящиеся к типам «изделия» и «компоненты», соответственно. Документ «изделие» можно разделить на логические части, а именно описание изделия и описание компонентов, входящих в состав изделия.

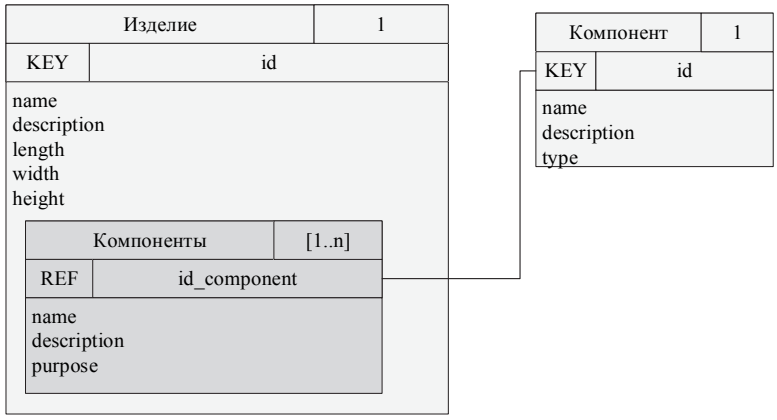

Рис. 1. Концептуальная модель
предметной области «изделие – компонент»

Описание сущности ПрО в виде нескольких частей документа позволяет указывать уникальные характеристики объекта в рамках одной коллекции. Свобода описания произвольной структуры документа дает преимущество в хранении полной информации, но способствует нарушению возможных ограничений ПрО. Таким образом, чтобы поддерживать корректность данных в БД, необходимы проверка нарушений правил ПрО и проверка семантических связей сущностей, участвующих в ассоциативных отношениях [5].

Расширенная документо-ориентированная модель данных с поддержкой ограничений целостности. Определим модель данных как множество, состоящее из объектов ПрО и ограничений целостности, накладываемых на них. Моделью называют объект, состоящий из двух множеств [6]:

$$M := < X, P >,$$

где X – непустое множество, являющееся носителем модели; P – множество предикатов (ограничений), заданных на множестве X.

Для описания и различия элементов, с помощью которых строится логическая модель документо-ориентированной базы данных, введен некоторый фиксированный алфавит V. Для указания факта, что элемент не принадлежит (или неизвестен в данный момент времени) алфавиту, вводится специальный символ Ω. Через V обозначим множество всех непустых слов в алфавите V^*. Тогда из этого множества V^* можно выделить следующие конечные множества.

Множество имен документов – O, в дальнейшем под именем документа будет пониматься сам документ.

Множество имен типов документов – $\boldsymbol{O}$, под именем типа документа будет пониматься сам тип документа.

Множество имен атрибутов документа – A, под именем атрибута будет пониматься атрибут.

Множество имен связей – R, под именем связи будет пониматься связь.

Множество имен типов связей – $\boldsymbol{R}$, под именем типа связи будет пониматься класс связи.

Множество имен свойств связей – RP, под именем свойства связи будет пониматься свойство связи.

Множество имен ограничений целостности – P, под именем ограничения будет пониматься само ограничение.

Пользователь базы данных отличает один объект от другого на основании своих представлений и знаний о предметной области и маркирует каждый объект уникальным именем. В предложенной модели данных уникальным именем выступает идентификатор, который представлен атрибутом A. Кроме того, отдельные экземпляры документов по некоторым критериям он объединяет в тип документа, например, если эти экземпляры обладают схожими (с некоторой вероятностью) наборами атрибутов. Каждый тип документа человек наделяет уникальным (в этой предметной области) именем (идентификатором) $o_i \in \boldsymbol{O}$, где o_i – конкретный экземпляр документа.

Основное отличие документа от атрибута с понятийной точки зрения является то, что внутренняя структура свойства документа не имеет значения и значение атрибута документа раскрывается только в рамках данного объекта.

Так же, как и документы, тип документа наделяется уникальным именем a_i ($a_i \in A$) (внутри документа).

Выделим в V^* конечное число множеств $D_1', ..., D_n'$ и обозначим через D_i множество $D_i' \cup \Omega$, $i = 1, ..., n$. Зафиксируем функцию N:

$$\forall a_i \in A \rightarrow N(a_i) = D_i,\ i = 1, ..., n,$$

где D_i будем называть доменом (набором допустимых значений) атрибута a_i, а множество всех пар $\{(a_i, d) \mid d \in D_i\}$ – атрибутом с именем a_i с множеством значений D_i. Пара (a_i, d) указывает, что атрибут с именем a_i имеет значение d.

Тогда формально можно описать документ *obj*:

$$obj := < o, < (a_1, d_1) \mid d_1 \in D_1), \ldots, (a_n, d_n) \mid d_n \in D_n) >,$$
$$p\ ((a_1, d_1), \ldots, (a_n, d_n)) >,$$

где *obj* – имя документа, что аналогично самому документу, так как под именем документа мы понимаем сам документ; $< (a_1, d_1) \mid d_1 \in D_1), ..., (a_n, d_n) \mid d_n \in D_n) >$ – набор атрибутов $a_i, ..., a_n$ (где $a_i \in A,\ i = 1, ..., n$) объекта со значениями $d_1, ..., d_n$,

соответственно; $p((a_1, d_1), \ldots, (a_n, d_n))$ – предикат над значениями атрибутов объекта $p \equiv p \in P$.

Такие документы будем считаться базовыми (реализация статической модели документа). Базовый документ представляется окрестностью нулевого порядка:

$$obj = \tau_o^{(0)},$$

Из базовых объектов с помощью связей могут быть сформированы более сложные объекты (реализация динамической модели документа).

Также верно:

$$\forall o \in O \rightarrow \exists \boldsymbol{O},$$

В общем случае понятие типа документа определяется через объекты, его составляющие:

$$obj = < \boldsymbol{o}, \{O_1 \ldots, O_n\} >,$$

где obj – тип документа, $\boldsymbol{o} \in \boldsymbol{O}$; $O_1 \ldots, O_n$ – множество документов, составляющих тип документов.

Как правило, два объекта, имеющие одинаковую структуру, относятся к одному типу документов. Рассмотрим их структуру:

$$obj_i := < o, < (a_1, d_1) \mid d_1 \in D_1), \ldots, (a_n, d_n) \mid d_n \in D_n) >,$$
$$p((a_1, d_1), \ldots, (a_n, d_n)) >$$
$$obj_j := < o, < (a_1, d_1) \mid d_1 \in D_1), \ldots, (a_n, d_n) \mid d_n \in D_n) >,$$
$$p((a_1, d_1), \ldots, (a_n, d_n)) >.$$

Упрощая описание структуры документа, получим, что объект obj можно выразить таким образом:

$$obj = < \boldsymbol{o}, \{(a_1), \ldots, (a_n)\}, p(a_i, \ldots, a_n) >,$$

где o – имя типа документа, $\boldsymbol{o} \in \boldsymbol{O}$; $a_1, \ldots, a_n$ – множество атрибутов для каждого документа такого типа объектов, $a_i \in A$, $i = 1, \ldots, n$; $p(a_1, \ldots, a_n)$ – структура ограничений значений свойств документов, входящих в тип.

В результате, вводя понятия вероятности появления атрибута в документе, получаем:

$$obj = < \boldsymbol{o}, \{(a_1, c_1), \ldots, (a_n, c_n)\}, p(a_i, \ldots, a_n) >,$$

где o – имя типа документа, $\boldsymbol{o} \in \boldsymbol{O}$; $a_1, \ldots, a_n$ – атрибуты документов, составляющих данный тип документа; $c_1, \ldots, c_n$ – вероятность наличия атрибута у произвольного объекта ‘o’ из класса ‘$\boldsymbol{o}$’, $c_i \in [0, 1]$.

На основании вероятностей $c_1, \ldots, c_n$ выводится неформальное понятие семантической близости. Под семантической близостью документов (или их частей) будем понимать функцию, характеризующую степень смысловой близости этих объектов (или их частей), которую задает проектировщик предметной области.

Семантическую близость можно выразить, например, значением, которое равно процентному соотношению инициализированных атрибутов ко всему набору атрибутов, каждый из которых имеет определенное значение, понятное проектировщику. Моделируя предметную область, проектировщик строит иерархию объектов предметной области, проходя от частного к общему и наоборот. Именно на этом этапе проектирования он относит различные по атрибутам объекты к одному базовому типу документов, используя понятие семантической близости.

Связи документов используются для создания составных документов из более простых. В этом случае простые документы могут группироваться в части внутри составного документа и выступать в роли атрибутов составного объекта. Часть, как было сказано выше, группирует какие-то свойства объекта по смыслу (семантике), назначению, принадлежности и т.д. Под связью r будет пониматься особый вид объектов следующего вида:

$$rel := < r, a_i, RP >,$$

где r – имя связи, $r \in R$; a_i – атрибут, участвующий в связи; RP – свойство связи, которое раскрывает его кратность.

Часть документа

$$k := < r_1, \ldots, r_i >,$$

где $r_1, \ldots, r_i$ – содержимое части – результаты связей.

Тогда любой составной документ q определяется посредством частей, его составляющих

$$q := < K, \{rp_q\} >,$$

где K – множество частей $\{k_1, ..., k_n\}$ из которых состоит объект; $\{rp_q\}$ – множество ограничений частей внутри объекта;

Документ называют *базовым*, если он имеет только одну часть $K = <k>$.

Документ называют *простым*, если он является базовым объектом, который может содержать атрибуты, не содержащиеся в базовом объекте, и состоит всего из одной части.

Документ называют *сложным*, если используемая в нем часть k_i является частью (частями) другого простого документа.

Приведенная модель является расширением документо-ориентированной модели с поддержкой ограничений ссылочной целостности и ограничений предметной области.

Выводы. Дано описание математической модели расширенной документо-ориентированной базы данных, определяющей ограничения ссылочной целостности и ограничения предметной области. Ограничения, накладываемые на уровне модели данных и реализуемые на уровне системы управления базой данных, гарантируют выполнение заданных пользователем ограничений предметной области для систем автоматизированного проектирования.

Литература

1. *Гатчин Ю.А., Донецкая Ю.В.* Разработка методик для автоматизации проектирования изделий приборостроения // Труды конгресса по интеллектуальным системам и информационным технологиям «AIS-AT'09»: в 4 т. М.: Физматлит, 2009. Т. 3. С. 145–149.

2. *Дехтярь М.И., Диковский А.Я., Спиратос Н.* Восстановление ограничений целостности за счет наименьших достаточных изменений // Программирование. 1998. № 2. С. 27–37.

3. *Лучинин З.С. Сидоркина И.Г.* Модуль поддержки ограничений целостности в документо-ориентированных базах данных // Информационные технологии в профессиональной деятельности и научной работе: материалы Всерос. науч.-практ. конф. Йошкар-Ола, 2014. С. 56–59.

4. *Лучинин З.С., Сидоркина И.Г.* Формализация семантики в документо-ориентированных базах данных // Вестник Поволжского государственного технологического университета. 2014. № 3. С. 57–65.

5. *Лучинин З.С., Сидоркина И.Г.* Формирование ссылочной целостности в документо-ориентированных базах данных // Информационные технологии в электронике и электроэнергетике: материалы 9-й Всерос. науч.-техн конф. Йошкар-Ола, 2014. С. 339–341.

6. *Новиков Ф.А.* Дискретная математика для программистов. 2-е изд. М.: Питер, 2007. С. 68.

7. *Тальхайм Б.* Обзор семантических ограничений для моделей баз данных [Электронный ресурс]. URL: http://www.intsys.msu.ru/magazine/archive/v3(3-4)/thalheim-307-351.pdf (дата обращения: 20.10.2014).

8. García-Solaco M., Saltor F., Castellanos M. Semantic heterogeneity in multidatabase systems. In: Bukhres O.A., Elmagarmid A.K., ed. Object-oriented multidatabase systems. Prentice Hall, 1996, pp. 129–195.

References

1. Gatchin Yu.A., Donetskaya Yu.V. *Razrabotka metodik dlya avtomatizatsii proektirovaniya izdelii priborostroeniya* [Development of techniques for design automation product engineering]. *Trudy kongressa po intellektual'nym sistemam i informatsionnym tekhnologiyam «AIS-AT'09»: v 4 tomakh* [Proceedings of the Congress on Intelligent Systems and Information Technology «AIS-AT'09». 4 vols]. Moscow, Fizmatlit Publ., 2009, vol. 3, pp. 145–149.

2. Dekhtyar' M.I., Dikovskii A.Ya., Spiratos N. *Vosstanovlenie ogranichenii tselostnosti za schet naimen'shikh dostatochnykh izmenenii* [Restoring integrity constraints due to the smallest sufficient changes]. *Programmirovanie* [Programming], 1998, no. 2, pp. 27–37.

3. Luchinin Z.S. Sidorkina I.G. *Modul' podderzhki ogranichenii tselostnosti v dokumento-orientirovannykh bazakh dannykh* [Integrity module of constraints in the document-oriented databases]. *Informatsionnye tekhnologii v professional'noi deyatel'nosti i nauchnoi rabote: materialy Vseros. nauch.-prakt. konf.* [Proc. of Russ. conf. «Information technologies in professional work and scientific work»]. Yoshkar-Ola, 2014, pp. 56–59.

4. Luchinin Z.S., Sidorkina I.G. *Formalizatsiya semantiki v dokumento-orientirovannykh bazakh dannykh* [Formalization of the semantics of the document-oriented databases]. *Vestnik Povolzhskogo gosudarstvennogo tekhnologicheskogo universiteta* [Bulletin of the Volga State Technological University], 2014, no. 3, pp. 57–65.

5. Luchinin Z.S., Sidorkina I.G. *Formirovanie ssylochnoi tselostnosti v dokumento-orientirovannykh bazakh dannykh* [Forming of referential integrity in the document-oriented databases]. *Informatsionnye tekhnologii v elektronike i elektroenergetike: materialy 9-i Vserossiyskoy nauch.-tekhn. konf.* [Proc of 9th Russ. Conf. «Information technologies in electronics and power»]. Yoshkar-Ola, 2014, pp. 339–341.

6. Novikov F.A. *Diskretnaya matematika dlya programmistov. 2-e izdanie* [Discrete mathematics for computer programmers: a textbook for high schools. 2nd ed.]. St. Petersburg, Piter Publ., 2007.

7. Tal'khaim B. *Obzor semanticheskikh ogranichenii dlya modelei baz dannykh* [Overview of semantic constraints for database models]. Available at: http://www.intsys.msu.ru/magazine/archive/v3(3-4)/thalheim-307-351.pdf (accessed 20 November 2014).

8. García-Solaco M., Saltor F., Castellanos M. Semantic heterogeneity in multidatabase systems. In: Bukhres O.A., Elmagarmid A.K., ed. Object-oriented multidatabase systems. Prentice Hall, 1996, pp. 129–195.

ЛУЧИНИН ЗАХАР СЕРГЕЕВИЧ – аспирант кафедры информационно-вычислительных систем, Поволжский государственный технологический университет, Россия, Йошкар-Ола (for.zahar@gmail.com).

LUCHININ ZAKHAR – post-graduate student of Computer Science Chair, Volga State University of Technology, Russia, Yoshkar-Ola.

СИДОРКИНА ИРИНА ГЕННАДЬЕВНА – доктор технических наук, профессор, декан факультета информатики и вычислительной техники, Поволжский государственный технологический университет, Россия, Йошкар-Ола (igs592000@mail.ru).

SIDORKINA IRINA – doctor of technical sciences, professor, dean of Informatics and Computer Engineering Faculty, Volga State University of Technology, Russia, Yoshkar-Ola.