Вестник УГАТУ
Уфа : УГАТУ, 2011 Т. 15, № 1 (41). С. 176–181

УДК 62-50

А. А. Левков

УПРАВЛЕНИЕ ДАННЫМИ В КОРПОРАТИВНЫХ ИС

Предлагается метод построения иерархических реляционных ИС по критерию семантического тождества атрибутов (алгоритмов), позволяющий производить полную нормализацию схемы данныхза небольшое количество шагов и реализовать дифференциальную модель описания алгоритмов. Использование данного метода существенно снижает избыточность схемы данных и алгоритмов и позволяет строить эффективные системы хранения и обработки данных. *База данных*; *реляционная модель*; *иерархическая реляционная модель*; *нормализация*; *семантическое тождество*; *расчет сложности*; *наследование*

ВВЕДЕНИЕ

Построение эффективной системы управления организационно-техническими объектами является одной из наиболее актуальных современных задач. Сложность построения таких систем обусловлена сложностью структуры и неоднородностью связей самих организационно-технических систем – они не поддаются строгой формализации, в их функционировании можно различить ряд параллельных потоков различной физической природы, зачастую неявно связанных друг с другом и подверженных внешним, часто непрогнозируемым воздействиям [1, 2].

В качестве средства автоматизации и поддержки принятия решений таких систем используются корпоративные информационные системы (КИС), реализованные на основе реляционных СУБД и направленные на интеграцию данных и обеспечение сквозного управления [1]. Следует отметить, что даже использование таких комплексных и объемных систем, как BAAN, SAP, IFS Applications не решает полностью данной проблемы – данные системы являются комплексами бизнес-приложений (их количество может доходить до 70), где каждый из модулей решает свои задачи. Даже при использовании полной сборки таких пост-ERP-систем трудно говорить об эффективной автоматизации – отдельные модули даже одной фирмы зачастую используют несовместимые форматы данных, требуют дублирования ввода информации – т. е. не образуют единой информационной среды. Общее количество сущностей в таких системах может превышать тысячи (в перспективе построение БД из десятков тысяч сущностей), что приводит к качественному росту сложности проектирования таких моделей [3, 4].

Существенным фактором, затрудняющим эксплуатацию КИС, является высокая динамика изменения схемы данных (СД) и алгоритмов их обработки, обусловленная изменениями как в самой реальной системе, так и в окружающем мире, и не имеющая отражения в средствах моделирования, разработки и поддержки КИС. В случае изменения СД КИС происходит «перепроектирование» только изменяющихся элементов схемы данных, без учета возможных воздействий вносимых изменений на всю КИС в целом.

Как при первичной разработке КИС, так и при ее развитии, наиболее острой проблемой построения СДКИС является нормализация реляционных отношений. Исходя из теории и практики нормализации, она осуществляется на выявленных функциональных зависимостях (ФЗ) между атрибутами в отношениях [5]. При выявлении ФЗ нельзя опираться на механический анализ уже существующих данных: если $r(R)$, $A \subseteq R$, $B \subseteq R$; $(A \to B)$, и $(\forall t_1, t_2 \in r: t_1(A) = t_2(A)) \Rightarrow (t_1(B) = t_2(B))$, так как сама природа КИС такова, что существенная часть работы системы заключается в сохранении *новой* информации. Т. е. существующая информация всегда неполна, и выявленные для нее функциональные зависимости могут оказаться неверными при поступлении новой информации [3].

Для проведения полной нормализации необходимо построение единого ненормализованного отношения с последующим проведением

$$O_{full} = \frac{|A|!}{(|A|-2)!} = |A|^2 - |A|$$

операций анализа данных на предмет определения функциональных зависимостей, где $|A|$ – количество атрибутов в ненормализованном отношении. Очевидно, что такие операции трудно вычислимы, так как количество атрибу-

Контактная информация: (347) 273-77-17

Работа выполнена при поддержке РФФИ, грант № 09-08-00490-а

тов в едином ненормализованном отношении может превышать десятки тысяч (при $|A| = 1000$ $O_{full} = 999000$) – столько экспертных операций анализа данных необходимо провести. Это вынуждает разработчиков осуществлять первичное разбиение модели на N отдельных сущностей интуитивно, и потом нормализовать эти, уже сравнительно небольшие структуры [6]. В таком случае необходимо выполнение $O_{part} = N(|A|^2 - |A|)$ операций анализа (если считать количество атрибутов в сущности одинаковым, то $O(CN)$), однако модель в таком случае нормализуется лишь частично и возможно возникновение «распыленных» сущностей – т. е. ситуаций, когда производные сущности, полученные при нормализации из разных первоначальных сущностей, являются отражениями одной реальной.

Также существенным недостатком современных КИС является отсутствие формализации алгоритмов обработки данных. Максимальный уровень связности для алгоритмов – использование шаблонов и макроопределений на уровне БД КИС. Альтернативой является введение среднего слоя трансляции данных (Linq и пр.), что позволяет использовать UML-модели алгоритмов, но приводит к «распаду» единой модели на отдельные, слабосвязанные модули. Незначительное изменение в СД КИС при этом может привести к существенным изменениям в алгоритмах обработки данных и наоборот [7].

Таким образом, для обеспечения управления данными в КИС в данной статье предлагается нормализованная иерархия типов в качестве основы построения СД КИС [8]. Данный подход может быть выражен в терминах реляционной алгебры и не требует дополнительных преобразований моделей. Он позволяет снизить семантическую избыточность СД КИС, упрощает операции по ее модификации.

Дифференциальное описание алгоритмов обработки данных на уровне БД, использование которых позволит снизить избыточность алгоритмического наполнения БД КИС, существенно упростить операции с данными и СД КИС и реализовать «самоконструирующиеся» системы.

Метод расчета описательной сложности СД КИС, позволяющий проводить их количественную оценку и сравнение, с целью выбора более эффективной СД КИС.

1.СТРУКТУРА ОРГАНИЗАЦИИ БД КИС

Для построения эффективной структуры СД КИС автор предлагает использовать критерий семантического тождества атрибутов реляционных отношений и производить иерархизацию СД путем слияния семантически тождественных атрибутов в отношениях в новые реляционные сущности, с последующим их связыванием по первичным ключам (построение иерархии типов/классов).

Если существуют отношения $X_1\{x_1, x_2,\ldots,x_n\}$ и $Y_1\{y_1, y_2,\ldots,y_n\}$, содержащие подмножества атрибутов $X'\{x'_1, x'_2,\ldots,x'_m\} \subset X$ и $Y'\{y'_1, y'_2,\ldots,y'_m\} \subset Y$, такие, что $\mathrm{Sem}(x_i) \equiv \mathrm{Sem}(y_i)$ (элементы семантически тождественны), где $x_i \in X'$, и первичные ключи отношений $PK_x \subset X'$ и $PK_y \subset Y'$ (это условие всегда выполняется для суррогатных ключей, либо напрямую, либо через функцию преобразования $f(PK, E)$), то формируются новые отношения:

$Z_1\{z_1, z_2,\ldots,z_m\}$, такое, что $\mathrm{Sem}(z_i) \equiv \mathrm{Sem}(x_i) \equiv \mathrm{Sem}(y_i)$,

$$Z = X' \cup Y',$$
$$X_1 = X - X',$$
$$Y_1 = Y - Y',$$

Отношения X_1 и Y_1 связываются вторичными ключами с отношением Z по первичным ключам $X_1.\mathrm{pk} \rightarrow Z.\mathrm{pk}$, $Y_1.\mathrm{pk} \rightarrow Z.\mathrm{pk}$, эти связи образуют иерархию, как показано на рис.1.

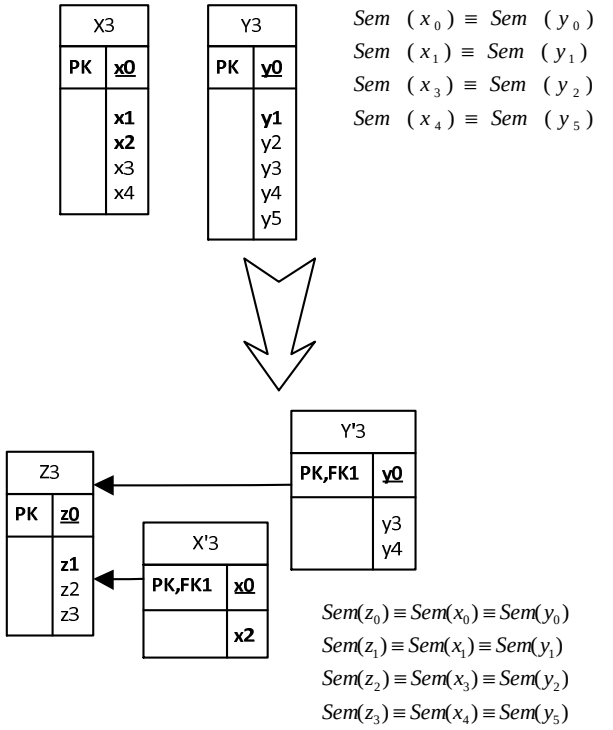

Рис. 1. Преобразование классической реляционной СД виерархическую

Таким образом, исходная сущность оказывается «распределенной» по иерархии новых сущностей, для обратного синтеза сущностей необходимо использовать рекуррентный алгоритм:

$$E^{full} = \begin{cases} \text{если } \nexists E_{\text{пред}}, \text{ то } E, \\ E \dfrac{pk}{\theta} E^{full}_{\text{пред}}, \end{cases}$$

$$A_E{}^{full} = \begin{cases} \text{если } \nexists E_{\text{пред}}, \text{ то } A_E, \\ A_E \cup A_{E^{full}_{\text{пред}}}. \end{cases}$$

Предлагаемый метод иерархизацииСД КИС позволяет существенно снизить ее семантическую избыточность и значительно упростить операции манипуляций со схемой данных [9].

2. НАСЛЕДОВАНИЕ АЛГОРИТМОВ

При формировании иерархии типов возникают дополнительные промежуточные сущности, ранее не существовавшие в классической реляционной СД. Такие сущности хранят обобщенные данные по дочерней группе сущностей. Их использование дает возможность реализовывать универсальные алгоритмы работы с данными и увеличивает скорость выполнения подобных запросов [10].

Если $M\{p_1, p_2,\ldots, p_n\}$ – множество алгоритмов, где каждый алгоритм определен как $p_i<S,M'>$, $i=\overline{1...n}$, $M' \subset M$ и S – множество элементарных обращений к данным (в случае БД – к атрибутам) $S<A'>$, где $A' \subset A\{a_1, a_2,\ldots, a_m\}$, то при преобразовании классической реляционной СД к иерархической вводятся новые атрибуты a_{new} по правилу:

$$Sem(a_{new}) = Sem(a_i) \equiv Sem(a_j), \quad i \neq \neq j,\ i = \overline{1...n},\ j = \overline{1...n} \qquad (1)$$

Таким образом, если в классической реляционной модели $\exists p_i$ и $\exists p_j$, такие, что $\forall Sem(a_k^i) \equiv Sem(a_k^j)$ и $M'^{\,i} = M'^{\,j}$, то $\text{Sem}(p_i) \equiv \text{Sem}(p_j)$ и с учетом (1) – введения новых атрибутов – в иерархической СД $p_i = p_j$, что позволяет определить в данном случае $M^{new} = M - p_j$. Т. е. происходит упрощение алгоритмического наполнения БД КИС путем слияния алгоритмов, обрабатывающих семантически подобные данные. Как показано на рис. 2.

Также, если в классической СД для s_k $\exists a_i$ и $\exists a_j$ такие, что $\text{Sem}(a_i) \equiv \text{Sem}(a_j)$, то с учетом (1) $s_k^{new} = s_k - a_j$. Т. е. упрощаются элементарные обращения к данным за счет уменьшения количества атрибутов в них.

Кроме того, построение иерархии сущностей делает возможным построение иерархии алгоритмов обработки данных в сущностях, т. е. использование дифференциального метода реализации алгоритмов – наследование алгоритмов.

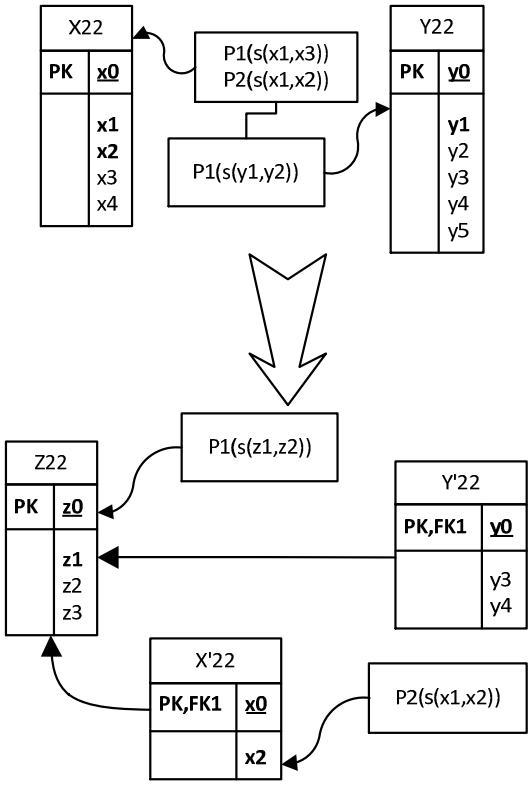

Рис. 2. Слияние алгоритмов в иерархической СД

Т.е. если есть некое поддерево наследования – множество сущностей $E(e_1, e_2,\ldots, e_n)$, где каждой сущности сопоставлены методы $M_i^{'}(M_i^{'} \subset M)$, $j=\overline{1...n}$, то на сущности e_i разрешено выполнение не только множества методов $M_i^{'}$, но и всех методов родительских сущностей, т.е. согласно реккурентному алгоритму:

$$M_i^{'+} = \begin{cases} \text{если } \nexists E_{\text{пред}}, \text{ то } M_i^{'}, \\ M_i^{'} + M_{\text{пред}}^{'+}. \end{cases}$$

Наследование алгоритмов позволяет существенно уменьшить алгоритмическую избыточность системы и повышает надежность функционирования всей системы.

3.РАСЧЕТ СЛОЖНОСТИ СД КИС

Автор предлагает оценивать структурную сложность реляционной СД через общее количество элементов каждого типа. Таким образом, структурная сложность может быть представлена как многомерный вектор, где композиционность выражается через добавление дополнительной координаты вектора, величина которой

равна мощности множества элементов данного типа (т. е., к примеру, если у нас есть 5 корзин по 10 яблок в каждой, то мы имеем не только 50 яблок, но и 5 корзин). Это можно выразить следующим образом:

$$S^Z = \sum_{j=1}^{|Z|} Z_j \left\langle |\, x_j \,|, |\, y_j \,| \right\rangle = \left\langle |\, Z \,|, \sum_{j=1}^{|Z|} |\, x_j \,|, \sum_{j=1}^{|Z|} |\, y_j \,| \right\rangle,$$

где Z – множество композиционных элементов Z_j; $j \in [1..|Z|]$ ($|Z|$ – мощность множества Z), где каждый Z_j определен на множествах x_j и y_j.

Структурно-алгоритмическая сложность СД (S^{DM}) является суперпозицией структурной (S^{DM}_{struct}) и алгоритмической (S^{DM}_{alg})сложностей:

$$S^{DM} = S^{DM}_{struct} + S^{DM}_{alg}. \qquad (2)$$

Формульные выражения структурной сложности можно представить следующим образом:

$$S^{DM}_{struct} = \left\langle S^E, S^I, |\, C \,|, S^V \right\rangle,$$

где S^E – сущностная сложность, S^I – индексная сложность, S^C – сложность связей, S^V – сложность представлений.

$$S^E = \left\langle |\, E \,|, \sum_{j=1}^{|E|} |\, A^{E_j} \,| \right\rangle, \qquad (3)$$

$$S^I = \left\langle |\, I \,|, \sum_{j=1}^{|E|} \sum_{k=1}^{I^{E_j}} |\, A^{I_k{}^{E_j}} \,| \right\rangle, \qquad (4)$$

$$S^V = \left\langle |\, V \,|, \sum_{j=1}^{|V|} |\, A^{V_j} \,|, \sum_{k=1}^{|V|} |\, O^{V_j} \,| \right\rangle. \qquad (5)$$

Алгоритмическая сложность для чисто реляционных систем может быть выражена следующим образом:

$$S^{DM}_{alg} = \left\langle |\, P \,|, S^{set}, S^{ins}, S^{upa}, S^{del} \right\rangle, \qquad (6)$$

где S^{sel} – сложность запросов на выборку, S^{ins} – сложность запросов на вставку, S^{upd} – сложность запросов на обновление, S^{del} – сложность запросов на удаление.

$$S^{sel} = \left\langle \sum_{j=1}^{|P|} \sum_{k=1}^{|sel^{P_j}|} |\, O^{sel_k^{P_j}} \; \sum_{j=1}^{|P|} \sum_{k=1}^{|sel^{P_j}|} |\, O^{sel_k^{P_j}} \,| \right\rangle, \qquad (7)$$

$$S^{ins} = \left\langle |\, ins \,|, \sum_{j=1}^{|P|} \sum_{k=1}^{|ins^{P_j}|} |\, A^{ins_k^{P_j}} \; \sum_{j=1}^{|P|} \sum_{k=1}^{|ins^{P_j}|} |\, O^{ins_k^{P_j}} \,| \right\rangle, \qquad (8)$$

$$S^{upd} = \left\langle |\, upd \,|, \sum_{j=1}^{|P|} \sum_{k=1}^{|upd^{P_j}|} |\, A^{upd_k^{P_j}} \; \sum_{j=1}^{|P|} \sum_{k=1}^{|upd^{P_j}|} |\, O^{upd_k^{P_j}} \,| \right\rangle, \qquad (9)$$

$$S^{del} = \left\langle |\, del \,|, \sum_{j=1}^{|P|} \sum_{k=1}^{|del^{P_j}|} |\, O^{del_k^{P_j}} \,| \right\rangle. \qquad (10)$$

Полученные композиционные выражения сложности возможно привести к единому числовому параметру, воспользовавшись весовыми коэффициентами: $S = \sum_{j=1}^{N} k_j s_j$, где k_j – значимость j-го компонента сложности, которая определяется экспертным путем.

Таким образом, при помощи заданных формул возможно осуществление количественного расчета первичной сложности СД, как в векторном, так и нормализованном виде. Это позволяет сравнивать различные СД еще на этапе их проектирования, дает возможность оптимизировать их.

4. ОЦЕНКА ЭФФЕКТИВНОСТИ ИЕРАРХИЧЕСКОЙ СД КИС

Если считать дерево типов СД КИС сбалансированным и принять коэффициент семантической избыточности равным во всей иерархии и выразить его через степень узла дерева k, то основными ее характеристиками будут следующие:

Количество уровней в иерархии:

$$U = \frac{\ln(N)}{\ln(k)} + 1.$$

Общее количество первичных элементов (сущностей) в иерархии:

$$|\, X' \,| = \frac{|\, X \,| k - 1}{k - 1} \approx |\, X \,| * \frac{k}{k - 1}.$$

Количество распределяемых (вторичных) по иерархии свойств сущности(атрибутов, алгоритмов, индексов, ссылок, представлений): $|\, X'_E \,| = \frac{|\, X \,|}{|\, E \,| U'}$, что при $|E|>>0$ позволяет определить как общее количествотаких свойств для СД КИС, как

$$|\, X'' \,| \approx |\, X' \,| * \frac{k}{U(k-1)}.$$

Данная зависимость имеет кумулятивный эффект виерархической СД КИС: количество производных элементов для распределяемых свойств (атрибутов в индексах, представлениях, операторов в алгоритмах)– третичных элементов – будет иметь вид

$$|\, X''' \,| \approx |\, X' \,| * \left(\frac{k}{U(k-1)} \right)^2.$$

На графиках ниже отображены зависимости относительной сложности ($|X'|$ / $|X|$) в иерархической СД КИС для каждого типа элемента.

Как можно видеть из представленных зависимостей, иерархическая СД КИС показывает рост эффективности с ростом количества элементов в структуре.

Исходя из зависимостей (2–10), общая описательная сложность иерархической СД КИС можно представить следующим образом:

$$S^{hier} = \{1\}X' + \{1\}X'' + \{14\}X''',$$

т. е. использование иерархической организации приводит к росту единственного параметра – количества сущностей, 4 параметра уменьшаются по закону X'', 14 – по закону X'''. Если принять веса всех компонентов сложности одинаковыми, то относительная сложность СД КИС (S^{DM}_{hier} / S^{DM}_{st}) выражается зависимостью, представленной на рис.4.

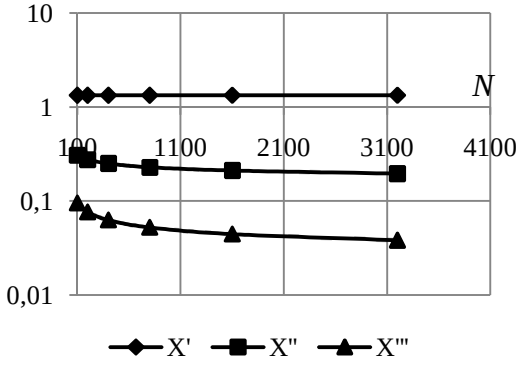

Рис. 3. Зависимость относительного количества элементов в иерархической СД КИС от количество первоначальных элементов (k = 4)

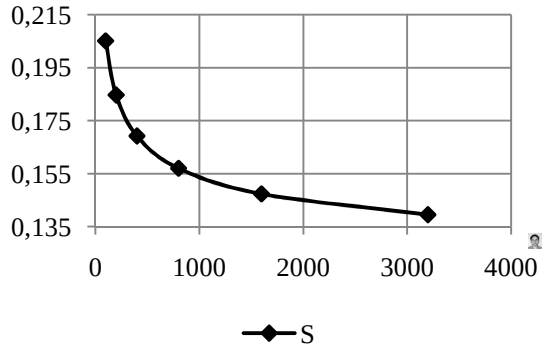

Рис. 4. Зависимость относительной взвешенной сложности иерархической СД КИС от количества сущностей в структуре (k = 4)

При преобразовании классической СД КИС в иерархическую для проведения полного анализа необходимо выполнить $O_{hier}^{rev} = \frac{|A|!}{2!(|A|-2)!} = \frac{|A|^2 - |A|}{2}$ автоматизированных операций, не требующих экспертного участия. Это связано с тем, что операция определения семантического тождества коммутативна, в отличие от функциональной зависимости.

При первичном построении иерархической СД КИС необходимо провести N экспертных разбиений на сущности, O_{hier}^{rev} автоматизированных операций по построению иерархии сущностей и $O_{hier}^{rev} = (N'(|A'|^2 - |A'|))$ операций по экспертному анализу функциональных зависимостей. Т. е. в иерархической СД КИС полная нормализация возможна за O_{hier}^{rev} экспертных операций (слияние семантически тождественных элементов делает невозможным появление «распыленных» сущностей)

На рис. 5 представлена зависимость относительной сложности нормализации иерархической и классической СД $k = O_{hier}^{prim} / O_{full}$ от количества сущностей в структуре.

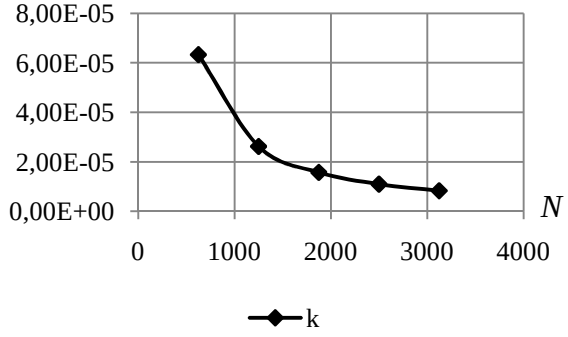

Рис. 5. Зависимость относительной сложности нормализации иерархической СД КИС от количества сущностей в структуре

Как видно из данной зависимости, проведение нормализации в иерархической СД КИС существенно проще, нежели в классической, и эффективность иерархической СД КИС увеличивается с ростом количества сущностей в системе.

ВЫВОДЫ

Построение эффективных СД КИС является важной научно-технической задачей. Классические методы нормализации структуры через единое ненормализованное отношение приводят к необходимости использовать $O(N^2)$ операций анализа в случае полной нормализации и $O(N)$ в случае частичной; отсутствие средств по описанию и структурированию алгоритмов не позволяет реализовать описание системы в рамках единого базиса.

В данной статье автором предлагается построение иерархической СД КИС, использование которой позволяет:

- Осуществить полную нормализацию СД за количество шагов $O(N / \ln(N))$, что позволя-

ет избежать избыточности в хранении данных;

- структурировать данные и алгоритмы в рамках единой структуры, что позволяет осуществлять эффективное управление ими;
- снизить семантическую избыточность данных и алгоритмов путем слияния семантически тождественных атрибутов на родительских уровнях иерархии, что приводит к существенному уменьшению сложности СД КИС.

Использование предлагаемой иерархической СД КИС позволяет управлять процессом построения и модификации структуры по критерию семантического тождества, а численный расчет сложности СД КИС позволяет производить сравнительный анализ различных моделей с целью выбора наиболее эффективной. Все это позволяет строить более эффективные СД КИС и автоматизированно производить их оперативную реконфигурацию.

СПИСОК ЛИТЕРАТУРЫ

1. IntersoftLab «Интеграция корпоративных приложений: основные понятия» // http://citcity.ru/11132/ (дата обращения: 25.08.2010).

2. **Гурьянов Л. В.** Интеграция АСУТП в АСУ предприятия // www.old.krug2000.ru/reports/08-ent_acs_integr.pdf (дата обращения: 25.08.2010).

3. **Khalilov A. I.** Data base organization in complex management information systems // Cybernetics and Systems Analysis, 02.02.2005.

4. **Cong Yu, Jagadish H. V.** Querying complex structured databases // VLDB. 2007.

5. **Codd E.F.** A relational model of data for large shared data banks // Communications of the ACM 13 .

6. **Chen P.** The entity-relationship model - toward a unified view of data // ACM Transactions on Database Systems (TODS), 1976.

7. **Багуи С.** Объектно-ориентированные базы данных: достижения и проблемы // Открытые системы. 2004. № 03.

8. Смит Д. М., Смит Д. К. Абстракции баз данных: агрегация и обобщение // СУБД. 1996. № 2.

9. **Кузнецов С.** Дубликаты, неопределенные значения, первичные и возможные ключи и другие экзотические прелести языка SQL // http://www.citforum.ru/database/articles/art_5.shtml (дата обращения: 25.08.2010).

10. **Rapaport M.** Object-Oriented Data Bases: The Next Step in DBMS Evolution // Comp. Lang. 1998. № 10.

ОБ АВТОРАХ

Левков Александр Александрович, доц. Дипл. магистр техники и технологий по информатике и вычисл. технике (УГАТУ, 2000). Канд. техн. наук по матем. и прогр. обеспечению вычисл. машин, комплексов и комп. сетей (УГАТУ, 2004). Иссл. в обл. баз данных, реляц. моделей данных, моделей физ. размещения данных.