УДК 004.891.3:004.658.2

А. Ю. Левченко,
А.С. Пригожев, канд.техн.наук

ИСПОЛЬЗОВАНИЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ ДЛЯ КЛАССИФИКАЦИИ КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Розглядаються питання побудови транзакційної моделі поведінки користувача корпоративної інформаційної системи (КІС). Для побудови такої моделі пропонується використовувати апарат теорії систем масового обслуговування – марковські моделі. Запропоновано критерій класифікації КІС на основі марковських моделей.

Рассматриваются вопросы построения транзакционной модели поведения пользователя корпоративной информационной системы (КИС). Для построения такой модели предлагается использовать аппарат теории систем массового обслуживания - марковские модели. Предложен критерий классификации КИС на основе марковских моделей.

The article represents the transaction model of corporate information system (CIS) user behaviour development. It is offered to use the scientific definition of the queuing theory – Markov models for such models construction. The CIS classification criteria based on the Markov models is offered.

Введение. Одним из способов повышения производительности корпоративных информационных систем (КИС), построенных на базе систем управления базами данных (СУБД), является подбор наилучших значений параметров конфигурации СУБД.

На производительность СУБД влияет набор различных факторов: структура запросов, вероятность появления конкретных запросов, интенсивность запросов к СУБД. При настройке параметров конфигурации важно учитывать все указанные факторы. Решение этой задачи затруднительно при ручной настройке параметров СУБД. Поэтому актуальной представляется задача автоматизации этого процесса.

Наиболее очевидным методом автоматизированной настройки параметров конфигурации является прямой перебор всех возможных значений параметров. Однако большинство параметров современных СУБД имеет большой диапазон значений, что в совокупности с их большим количеством делает такой алгоритм решения весьма трудоёмким с точки зрения временной сложности.

Для решения проблемы временной сложности автоматизированной настройки предлагается проводить классификацию КИС на основе транзакционной модели поведения пользователя – характера выполняемых им запросов и возможной последовательности их исполнения.

Как основу для предлагаемой классификации КИС можно использовать методики тестирования TPC (Transaction Processing Performance Council), включающие тестовые пакеты для оценки систем обработки транзакций и баз, данных разного типа и назначения [4]. Классы КИС, построенные на основе тестов TPC, в дальнейшем будем называть эталонными. Зная класс исследуемой системы и экспертные рекомендации для значений параметров настройки СУБД, можно значительно сократить количество и диапазоны значений тестируемых параметров конфигурации, что позволит сократить время поиска наилучших значений параметров и ускорит процесс принятия решения.

Модель транзакционного поведения пользователя. Для формализации структуры и связей БД можно воспользоваться фундаментальными понятиями реляционной алгебры. Для описания поведения пользовате-

ля предложено расширить данную алгебру введением в неё операционного профиля [3], который специфицирует использование СУБД в терминах операций (запросов) в системе и вероятностей их появления:

$$S = \langle DB, Q, \{СИГН\}, B \rangle, \quad (1)$$

где DB – структура базы данных, включающая подмножество отношений, атрибутов и связей между отношениями; $Q = \{q_1, q_2, q_3, ..., q_k\}$ – множество запросов, отсылаемых СУБД прикладным ПО, (k – количество запросов во множестве); СИГН – сигнатура, состоящая из объединения операций реляционной алгебры и операций вычисления вероятностей запроса к БД; B – операционный профиль.

Существующего описания операционных профилей пользователей как вероятностей появления операций в системе недостаточно для формального описания поведения пользователей. Поэтому предлагается использовать как операционный профиль марковские модели. Это позволит описать вероятности возможных последовательностей запросов в системе, что более полно отражает процессы, происходящие в реальной КИС, в частности, поведение клиентов КИС. Так как большинство известных типов КИС в немалой степени характеризуются поведением пользователей, представляется актуальным использование марковских моделей в качестве характеристики некоторого класса КИС.

При использовании марковских моделей для описания поведения пользователя, прежде всего, необходимо определить множество состояний модели.

Для множества Q справедливо следующее характеристическое свойство:

$$\forall q_i, q_j \in Q \,\&\, i \neq j : q_i \neq q_j, i, j \in \overline{1;k} \quad (2)$$

В (2) равенство запросов подразумевает как совпадение реляционных операций, так и совпадение имён таблиц и атрибутов в запросе.

Введём в рассмотрение временной интервал $[T_{min}; T_{max}]$, на котором зафиксировано множество моментов времени $T = \{t_1, t_2, t_3, ..., t_s\}$ получения СУБД запросов из множества Q (s – количество элементов множества T).

Также введём в рассмотрение множество U пользователей КИС $U = \{id_1, id_2, id_3, ..., id_p\}$, где id_p – идентификатор пользователя КИС; p – количество пользователей этой системы.

Введём подмножество декартова произведения:

$$L \subseteq T \times U \times Q. \quad (3)$$

Элементами множества, определяемого формулой (3) будут являться тройки l такие, что для них будет истинен характеристический предикат $EXECQ(t_i, id_j, q_l)$. Данный предикат принимает значение «истина» только в случае истинности утверждения: в момент времени $t_i \in T$, $i = \overline{1;s}$ клиентом $id_j \in U$, $j = \overline{1;p}$ был выполнен запрос $q_l \in Q$, $l = \overline{1;k}$.

Вводя обозначение $l = (t, id, q)$, формально множество L можно представить следующим образом:

$$L = \{ l \mid \forall l : EXECQ(l) = TRUE \}. \quad (4)$$

Зададим на множестве L отношение порядка между элементами

$$t_i < t_j \,\&\, i \neq j \Rightarrow l_i < l_j \quad (5)$$

где $l_i = (t_i, id, q)$, $l_j = (t_j, id', q')$, $i, j = \overline{1...s}$, $q, q' \in Q$, $id, id' \in U$.

В реальной СУБД множеству L, определяемому согласно (4), с заданным на нём отношением порядка (5) соответствует некоторый журнал запросов, который представляет собой записи о выполненных клиентом КИС запросах и времени их исполнения, упорядоченные в порядке их поступления в СУБД.

Введём в рассмотрение разбиение множества Q на m подмножеств, таких, что выполняются следующие соотношения:

$$Q = \bigcup_{i=1}^{m} Q_i \quad (6)$$

$$\bigcap_{i=1}^{m} Q_i = \varnothing. \quad (7)$$

Подмножества, для которых выполняются свойства (6) и (7), будем называть классами запросов.

Для разбиения используется следующее характеристическое свойство: в подмножество Q_i ($i = \overline{1;m}$) входят те $q \in Q$, которые

отличаются только наименованием таблиц и атрибутов таблиц, и не различаются по набору и порядку следования реляционных операций, а также условиям. Из формул (6) и (7) видно, что каждый элемент $q \in Q$ принадлежит только одному из подмножеств Q_i. Множество всех Q_i будем обозначать $\overline{Q}$.

Введём в рассмотрение множество шаблонов запросов:

$$S = \{ s_1, s_2, s_3, ..., s_n \}, \qquad (8)$$

где n – количество различных шаблонов.

Под шаблоном запроса будем понимать обычный SQL–запрос, в котором вместо имён таблиц и атрибутов используются некоторые переменные, которые могут принимать значения имён таблиц либо атрибутов в зависимости от их местоположения в шаблоне.

Построение некоторого шаблона запроса осуществляется на основе класса запроса, определяемого формулами (6) и (7). Данная операция осуществляется в два этапа: построение дерева разбора запроса [3] и замена в дереве реляционных операций имён таблиц и атрибутов на переменные.

Использование шаблонов запроса позволяет в значительной степени абстрагироваться от смыслового значения имён таблиц и объединять запросы с одинаковой структурой реляционных операций. Поэтому состояниями предлагаемой марковской модели будут шаблоны запросов (8), представленные в виде деревьев реляционных операций.

Введём взаимно-однозначное соответствие, $f : \overline{Q} \to S$ которое сопоставляет класс запросов $Q_i \in \overline{Q}$ с некоторым шаблоном $s_i \in S$.

Будем говорить, что существует переход между состояниями $s_i \in S$ и $s_j \in S$ в марковской модели поведения пользователя КИС в том случае, если для $(t_{i-1}, id, q), (t_i, id', q') \in L$ выполняются условия

$$q \in Q', q' \in Q''; Q', Q'' \in \overline{Q}, \qquad (9)$$

где $i = \overline{1...k}$.

Каждому переходу в марковской модели сопоставляется вероятность этого перехода, рассчитываемая как отношение:

$$P(s_i) = \frac{m_{si}}{m_{as}}. \qquad (10)$$

где m_{si} – количество переходов из состояния s_i в состояние s_j; m_{as} – общее количество переходов из состояния s_i $(i, j = \overline{1;n})$.

Заметим, что если для двух троек $(t_{i-1}, id, q_i) \in L$ и $(t_i, id', q_j) \in L$ выполняется условие (9) и при построении модели не учитываются компоненты $id.id' \in U$, то полученная марковская модель характеризует всю СУБД. Если же условие перехода (9) расширить дополнительным требованием $id = id'$, то в этом случае образуется множество моделей клиентов КИС.

Построенную таким образом марковскую модель будем называть марковской транзакционной моделью поведения пользователя.

При программной реализации модели используется журнал запросов СУБД, который состоит из элементов множества L.

Для классификации КИС на основе тестов ТРС необходимо решить вопрос о подобии транзакционных марковских моделей поведения пользователей теста и КИС в предположении, что тестирование на основе этих моделей даст приблизительно одинаковые результаты. Рассмотрим критерии такого подобия.

Подобие транзакционных марковских моделей поведения пользователя. Транзакционная марковская модель поведения пользователя характеризуется некоторым взвешенным графом, вершинами которого являются состояния – шаблоны запросов, дуги означают наличие переходов между состояниями модели и каждой дуге сопоставлена вероятность этого перехода.

Таким образом, чтобы характеризовать подобие марковских моделей, необходимо выявить значение трёх характеристик: количества совпадающих состояний у модели, количества и направления переходов в модели и пропорциональности вероятностей этих переходов.

Каждый эталонный класс КИС характеризуется своим множеством шаблонов запросов S_{ec} и транзакционной моделью пользователя, которая задаётся в виде матрицы

вероятностей переходов состояний M_{ec}. Соответственно, КИС характеризуется множеством шаблонов запросов S_{cis} и матрицей переходов состояния M_{cis}. Для характеристики степени совпадения запросов эталонного класса и КИС введём коэффициент подобия состояний, вычисляемый по формуле

$$k_1 = \frac{N_S}{N}, \qquad (11)$$

где N – мощность множества S_{ec}; N_S - мощность множества шаблонов запросов S, определяемого как $S = S_{ec} \bigcap S_{cis}$, где S_{ec} - множество шаблонов запросов эталонного класса КИС; S_{cis} - множество шаблонов запросов КИС.

Для анализа подобия графов воспользуемся тем свойством, что граф ориентированный. Построим матрицу смежности для графа эталонного класса B_{ec} согласно следующему закону:

$$b_{ij} = \begin{cases} 1, \textit{если } p_{ij} > 0 \\ 0, \textit{в противном случае} \end{cases}, \qquad (12)$$

где b_{ij} - элемент матрицы B_{ec}; p_{ij} - элемент матрицы M_{ec}.

Аналогично строится и матрица B_{cis}, представляющая собой матрицу смежности для марковской модели КИС. Матрицу B определим следующим образом

$$B = B_{ec} \oplus B_{cis}, \qquad (13)$$

где B_{ec}, B_{cis} – матрицы, построенные согласно (12).

Таким образом, матрица B является результатом побитового сложения «по модулю 2» матриц B_{ec} и B_{cis}. При вычислении матрицы B (см. формулу (13) возможны три варианта:

1. Размерности матриц B_{ec} и B_{cis} совпадают. В этом варианте сразу проводятся вычисления по формуле (13).

2. Размерность матрицы B_{ec} больше размерности матрицы B_{cis}. Перед вычислением по формуле (13) матрица B_{cis} расширяется нулевыми строками и столбцами, соответствующими состояниям, которые есть в B_{ec}, но нет в B_{cis}.

3. Размерность матрицы B_{cis} больше размерности матрицы B_{ec}. Тогда, перед вычислением по формуле (13), матрица B_{ec} расширяется нулевыми строками и столбцами. Они будут соответствовать состояниям, присутствующим в B_{cis}, но отсутствующим в B_{ec}.

Введём коэффициент

$$k_2 = \frac{N_0}{M_s^2}, \qquad (14)$$

где N_0 – количество нулевых элементов в матрице B; M_s – мощность множества S.

Исходя из определения операции «сложения по модулю 2», можно говорить о том, что коэффициент k_2 указывает на степень подобия структур графов марковских моделей и КИС.

Для оценки подобия марковских моделей и КИС необходимо также, чтобы вероятности переходов данных графов можно было аппроксимировать с помощью регрессии. Для оценки этого требования рассмотрим вектора вероятностей X_{ec} и X_{cis}, сопоставленных соответственно рёбрам графа эталонного класса и КИС. Данные вектора можно получить из матриц M_{ec} и M_{cis} путём их обхода по одному и тому же закону. Отметим, что в случае несовпадения размерностей матриц необходимо выполнить расширение соответствующих матриц нулевыми строками и столбцами согласно правилам 2 и 3, описанным выше.

Далее, для построенных векторов рассчитаем коэффициент корреляции[2]

$$k_3 = \left| \frac{cov(X_{ec}, X_{cis})}{\sqrt{DX_{ec} * DX_{cis}}} \right|, \qquad (15)$$

где $cov(X_{ec}, X_{cis})$ – ковариация векторов X_{ec} и X_{cis}; DX_{ec} – дисперсия вектора эталонного класса; DX_{cis} – дисперсия вектора матрицы КИС.

Для нашего случая характер зависимости (прямая или обратная) не важен, поэтому в формуле (15) содержится знак модуля.

Из определения коэффициента корреляции и формулы (15) следует, что $0 \le k_2 \le 1$.

Для оценки статистической значимости коэффициента корреляции воспользуемся критерием Стьюдента

$$\frac{k_3\sqrt{n-2}}{\sqrt{1-k_3^2}} \le t_{1-\frac{\alpha}{2}}(n-2), \qquad (16)$$

где $n-$ размерность векторов X_{ec} и X_{cis}; $t_{1-\frac{\alpha}{2}}(n-2)$ – α-квантиль распределения Стьюдента.

Если неравенство (16) выполняется, то k_3 принимается равным 0, что соответствует независимости векторов X_{ec} и X_{cis}. В противном случае для дальнейшей оценки значение k_3 принимается равным значению, вычисленному по формуле (15) с доверительной вероятностью $1-\frac{\alpha}{2}$.

Используя вычисленные ранее коэффициенты, определим интегральную оценку принадлежности КИС некоторому классу:

$$k = \sum_{i=1}^{3} w_i k_i , \qquad (17)$$

где w_i – весовой коэффициент, который выбирается с учётом мнений экспертов; k_i – значения, вычисленные по формулам (11), (14) и (15) соответственно.

Использование предложенного критерия оценки подобия транзакционных марковских моделей поведения пользователя позволяет сократить время автоматизированной настройки параметров СУБД. Для этого каждому эталонному классу сопоставляется набор оптимальных параметров настройки. При значениях коэффициента k, близких к 1, данный набор параметров применяется к СУБД. В противном случае определяются два эталонных класса, для которых значение k максимально. Значения параметров, сопоставленных этим классам, дают границы интервалов, в которых производится нагрузочное тестирование СУБД с целью выявления оптимальных параметров настройки.

Формализация базы данных Одесского национального политехнического университета (ОНПУ) в терминах транзакционной поведенческой модели пользователя

Приведем пример описания структуры БД реально существующей КИС предназначенной для учета кадров Одесского национального политехнического университета(ОНПУ). Для описания физической структуры БД представим отношения (таблицы) в виде множества двоек вида

$$\mathrm{R} = \left(\langle a_1 : d_1\rangle, \langle a_2 : d_2\rangle, ..., \langle a_n : d_n\rangle\right), \qquad (17)$$

где a_n – атрибут отношения для описания столбцов таблиц; d_n – домен, определяющий тип данных заданного атрибута.

Выполним унификацию таблицы «личность», включающую следующие атрибуты: номер (числовой), фамилия (текст), имя (текст), отчество (текст), пол (символ), идентномер(текст).

rel1 = (<atr1:int>, <atr2:varchar>, <atr3:varchar>, <atr4:varchar>, <atr5:char>, <atr6:char:PK>)

Таблица «сведработа» включает атрибуты: номер (числовой), кодподразд (числовой), коддолжность (числовой), датувольн (дата), кодтипработы (числовой), ставка (десятичное), деньги (текст), кодсостав (числовой), датперемещения (дата), кодкатегор (числовой), оклад (десятичное), аванс (десятичное), кодначуд (числовой), коднеоблмин (числовой), шифрзатрат (текст), датвступдолжн (дата), примечание (текст), кодставки (числовой), флагоснместраб (символ), отсутствует (символ), замещен (символ), табномер (числовой), приказномер (текст), приказдата (дата), бухгалтерия (числовой), принят_на_работу (дата), ранг (числовой), разряд (числовой).

rel2 = (<atr1:int:PK>, <atr2:int:PK>, <atr3:int:PK>, <atr4:date>, <atr5:int>, <atr6:double>, <atr7:char>, <atr8:int>, <atr9:date>, <atr10:int>, <atr11:double>, <atr12:double>, <atr13:char>, <atr14:smallint>, <atr15:char>, <atr16:date>, <atr17:varchar>, <atr18:int>, <atr19:char>, <atr20:char>, <atr21:char>, <atr22:int>, <atr23:varchar>, <atr24:date>, <atr25:smallint>, <atr26:date>, <atr27:smallint>, <atr28:smallint>)

Таблица «подразделения» включает атрибуты: кодподразд (числовой), назвподр (текст), сокрназвподр (текст), кодродитподр (числовой), состав (текст), статус (текст), приоритет (числовой), назвродитпадеж

(текст).

rel3 = (<atr1:int:PK>, <atr2:varchar>, <atr3: varchar >, <atr4:int>, <atr5:char>, <atr6:varchar>, <atr7:int>, <atr8:varchar>)

Таблица «должность» включает атрибуты: коддолжность (числовой), назвдолжность (текст), сокрназвдолжность (текст), приоритет (числовой), кодсостав (числоой).

rel4 = (<atr1:int:PK>, <atr2:varchar>, <atr3:varchar>, <atr4:int>, <atr5:int>)

Рассмотрим пример построения дерева запроса для анализируемой КИС на основе ее журнала транзакций. Выделяем шаблоны запросов по запросам лог файла. В результате получаем множество шаблонов – наших состояний. Возьмем произвольный запрос:

>>> 2008-09-10 11:00:30 EEST||10.1.6.19(56463)||u4arm||11315||
<<<LOG: statement: select b.oid, a.фамилия , a.имя, a.отчество, c.назвродпадеж, d.назвродитпадеж, a.пол, b.ставка, b.деньги, b.оклад, b.кодсчет, b.коддолжность, b.кодподразд

from сведработа b, личность a, подразделения d, должность c

where a.номер = 3466

and a.номер = b.номер

and b.кодподразд = d.кодподразд

and b.коддолжность = c.коддолжность

and b.кодсостав = 1

and b.кодподразд = 45

and b.датувольн = '30.06.2008'

and b.кодтипработы in (8)

and (not exists (select * from сведработа where кодсостав = 1 and кодтипработы in (8) and кодподразд = 45 and датувольн >= '1.09.2008' and b.номер = номер))

Преобразуем запрос в соответствии с формулами (5) и (6):

L = (2008-09-10 11:00:30, 10.1.6.19(56463), «select …»)

Далее выполняем унификацию запроса типа SWF (“select-from-where”), т.е. заменяем имена таблиц и атрибутов на переменные, определенные ранее в процессе унификации таблиц:

<SWF>::=SELECT <SelList1> FROM <FromList1> WHERE <Condition1>

<SelList1>::= rel2.oid, rel1.atr2, rel1.atr3, rel1.atr4, rel3.atr8, rel4.atr3, rel1.atr5, rel2.atr6, rel2.atr7, rel2.atr11, rel2.atr5, rel2.atr3, rel2.atr2; <FromList1>::= rel1, rel2, rel3, rel4;

<Condition1>::= <rel1.atr1 = (int) value> AND <rel1.atr1 = rel2.atr1> AND <rel2.atr2 = rel3.atr1> AND <rel2.atr3 = rel4.atr1> AND <rel2.atr8 = (int) value> AND <rel2.atr2 = int> AND <rel2.atr4 = ’date’> AND <rel2.atr5 in ((int) value)>

AND <(NOT EXISTS (<SWF>::= SELECT <SelList2> FROM <FromList2> WHERE <Condition2>)>

<SelList2>::= *;

<FromList2>::= rel2;

<Condition2>::= <rel2.atr8=int> AND <rel2.atr5 in (int) value> AND <rel2.atr2 = (int) value> AND <rel2.atr4>=’date’> AND <rel2.atr1=(int)value>.

Согласно стандарту SQL первым выполняется вложенный запрос SELECT * FROM rel2 WHERE …, в результате которого вначале выполняется операция выборки из таблицы, а затем операция исключения (проекции) группы записей, удовлетворяющих заданным условиям.

Для приведенных запросов можно синтезировать следующее дерево разбора [1], приведенное на рисунке.

Выводы. В результате проведённого исследования получил дальнейшее развитие аппарат марковских моделей для использования его при решении задачи моделирования поведения пользователя КИС. Предложено в качестве состояний модели использовать шаблон запроса, представляющий собой описание последовательности реляционных операций в запросе и условий этих операций. Это позволяет сократить количество состояний модели.

Предложена классификация КИС на основе тестов TPC. Основной характеристикой класса является транзакционная марковская модель поведения пользователя. Каждый класс КИС сопоставляется с набором оптимальных параметров настройки.

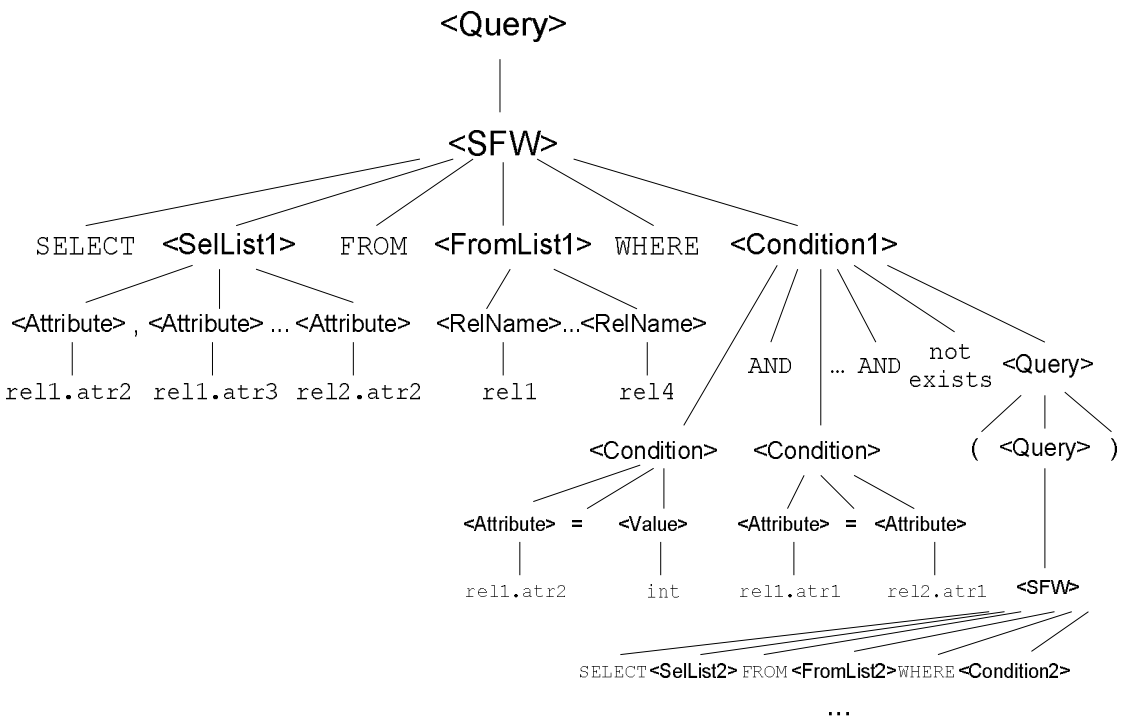


Рисунок. Дерево разбора

На основании транзакционной марковской модели поведения пользователя предложен критерий подобия моделей, применяя который можно проводить автоматизированную настройку параметров СУБД либо значительно сокращать количество и интервалы значений параметров для настройки модели, за счёт сопоставления описанию класса оптимального набора параметров.

Предложенные подходы позволяют синтезировать в дальнейшем экспертную систему настройки параметров СУБД. Её использование администратором КИС позволит сократить время подбора оптимальных параметров настройки.

Список использованной литературы

1. Гарсиа-Молина Г. Системы баз данных /Гарсиа-Молина Г., Ульман Дж., Видом Дж.. Пер. с англ. – М.: Изд.дом «Вильямс», 2003. – 1088 с.: ил.

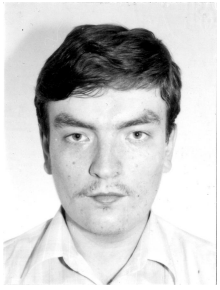
2. Горяинов В.Б. Математическая статистика: Учеб. для вузов / В.Б. Горяинов, И.В. Павлов, Г.М. Цветкова и др.; под ред. В.С. Зарубина, А.П. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана. – 2001. – 424 с.

3. Коваль Г.И. Концепция профилей в инженерии надежности программных систем /Коваль Г.И., Мороз Г.Б., Коротун Т.М. // Математичні машини і системи. – 2004. – № 1. – С. 166–184.

4. O. Serlin. The History of DebitCredit and the TPC. The Benchmark Handbook for Database and Transaction Processing Systems, Jim Gray (Ed.), Morgan Kaufmann, 2nd Ed. 1993, pp. 21-40.

Получено 15.03.2010

Левченко Александра Юрьевна, ассистент каф. системного программ.обеспечения Одесск. нац. политехн. ун-та, 7348-566



Пригожев Александр Сергеевич, канд. техн. наук, ст. преп. каф.системного программного обеспечения Одесск.нац. политехн.ун-та, 7348-566