

ПОБУДОВА КЛАСІВ SQL-ЗАПИТІВ ДЛЯ МОДЕЛЮВАННЯ ПОВЕДІНКИ КОРИСТУВАЧІВ ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕСТУВАННЯ ЇХ ПРОДУКТИВНОСТІ

Вступ. Одним з етапів життєвого циклу інформаційних систем (ІС) є стадія розвитку, яка характеризується збільшенням кількості користувачів і об'ємів інформації, що зберігається. У зв'язку з цим продуктивність ІС може знижуватися. Підвищення ефективності функціонування окремих компонентів ІС, наприклад, системи керування базами даних (СКБД), дозволяє підвищити продуктивність всієї системи в цілому. Процес підвищення продуктивності включає етап аналізу ІС, тобто тестування. Тестування може виконуватися з використанням стандартних методик тестування або спеціалізованих тестів продуктивності. Стандартні тести не завжди дозволяють симулювати реальну взаємодію користувача з ІС, що впливає на достовірність результатів тестування і знижує ефективність заходів щодо поліпшення характеристик продуктивності ІС. Проте спеціалізовані тести, що будуються на основі структури досліджуваної бази даних (БД) і запитів користувачів ІС, підвищують репрезентативність результатів тестування. Таким чином, актуальною є проблема побудови зазначених тестів.

Аналіз останніх досліджень і публікацій. Вихідними даними для побудови тесту є модель транзакційної поведінки користувача [1, 2], яка характеризує поведінку користувача ІС на основі SQL-запитів, що він надсилає до БД. Модель описує систему за допомогою зваженого орієнтованого графа (рис. 1), вершинами якого є шаблони SQL-запитів $s_i \in S$, а дуги та їх спрямування вказують на послідовність виникнення запитів, тобто порядок надходження запитів в ході звернення користувача ІС до БД. Вага дуги $P(s_i)$ характеризує частоту переходів між шаблонами запитів користувачів.

Для побудови шаблону SQL запиту у первинному запиті (рис. 2) виділяються основні SQL конструкції типу SELECT, FROM, WHERE. Імена таблиць і атрибутів замінюються на деякі змінні. Це дозволяє абстрагуватися від смислового навантаження запиту, сконцентрувавши увагу на його конструкції.

Джерелом для побудови множини S шаблонів запитів користувачів виступає журнал транзакцій, який являє собою записи про виконані користувачами БД ІС запити і час їх виконання, які упорядковані в порядку їх надходження до СУБД. Слід зазначити, що множина шаблонів SQL-запитів, яка отримана шляхом простого перетворення з SQL-запитів, може бути багатоваріантною, а, отже, надлишковою для зберігання моделі транзакційної поведінки користувача і подальшої побудови тестового середовища.

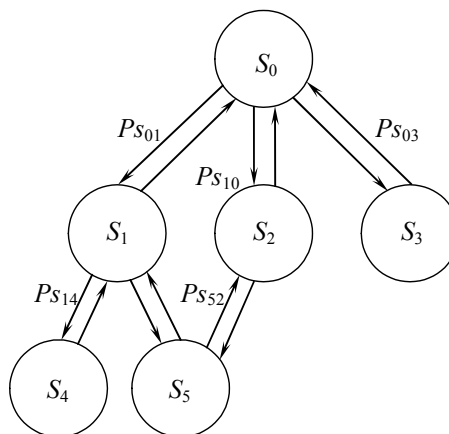


Рис. 1. Граф поведінки користувача ІС

Метою дослідження є скорочення варіативності досліджуваної вибірки шаблонів запитів за рахунок класифікації шаблонів SQL-запитів користувачів, яка дозволить зберігати не сам шаблон, а його клас.

SELECT

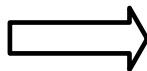
b.oid, a.фамилия, a.имя, a.отчество,
с.назвродпадеж, d.назвродитпадеж, a.пол,
b.ставка, b.деньги, b.оклад, b.кодсмет,
b.коддолжность, b.кодподразд

FROM

сведработа b, личность a, подразделения d,
должность с

WHERE

a.номер = 3466
and a.номер = b.номер
and b.кодподразд = d.кодподразд
and b.коддолжность = с.коддолжность
and b.кодсостав = 1
and b.кодподразд = 45
and b.датуволен = '30.06.2012'
and b.кодтипроботы in (8)



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)>

Рис. 2. Приклад побудови шаблону SQL-запиту

Викладення основного матеріалу. Завдання класифікації шаблонів запитів зводиться до присвоєння булевого значення кожній парі

$$(s_i, c_j) \in Q \times C,$$

де S — множина шаблонів запитів;

$C = \{c_1, \dots, c_n\}$ — множина визначених категорій.

Таким чином, цільова функція класифікації

$$\Phi: Q \times C \rightarrow \{T, F\},$$

де T — “істина”;

F — “хибність”.

Для віднесення шаблону запиту до того чи іншого класу необхідно попередньо визначити основу класифікації, тобто сформуванню множини еталонних класів шаблонів запитів.

Доцільно сформуванню еталонну множину шаблонів запитів, використовуючи існуючі промислові стандарти, такі, як методики тестування TPC (Transaction Processing Performance Council) [3...6], які описують актуальні на сьогоднішній день типи систем обробки транзакцій. Стандарти TPC-C, TPC-H, TPC-E (OLTP) включають транзакції з простими SQL-запитами до БД (додавання даних, видалення даних з угрупованням, з об'єднанням), а стандарт TPC-D (DSS) включає транзакції зі складними аналітичними запитами.

Множину еталонних шаблонів запитів з транзакцій TPC необхідно розбити на підмножини, які описують різні типи SQL запитів. Дані підмножини надалі будуть описувати еталонні класи шаблонів SQL-запитів.

Для формування еталонів шаблонів запитів розроблено алгоритм, який включає наступні етапи:

- виокремлення множини SQL-запитів з транзакцій тестів TPC;
- синтаксичний розбір SQL-запиту;
- семантичний аналіз SQL-запиту;
- побудова шаблону SQL-запиту;
- кластеризація множини шаблонів запитів;
- збереження результатів у вигляді еталонних класів шаблонів запитів.

Кластеризація отриманих на передостанньому кроці шаблонів виконується з метою визначення кластерів (підмножин запитів) зі схожими властивостями, тобто класів, та формування множини еталонних шаблонів, що їх описують.

Обираючи метод кластеризації слід враховувати необхідність зберігати смислову характеристику запиту. Аналіз показав доцільність використання методу ієрархічної кластеризації [7], який не вимагає заздалегідь визначати число кластерів і задавати відстань між їх центрами.

Ієрархічний підхід кластеризації будує систему вкладених розбиттів (кластерів). Таким чином, на виході алгоритму отримуємо дерево кластерів, коренем якого є вся вибірка, а листям — дрібні кластери.

Характеристиками вектора кластеризації виступають стандартні конструкції та операції мови SQL, такі як:

- операції з'єднання ($\{ \text{INNER} \mid \text{LEFT} \mid \text{RIGHT} \} \text{ JOIN}$);
- операції порівняння (" $=$ " " $<$," " $>$," " $\leq$ " " $\geq$ " або " $<>$.");
- агрегатні функції (SUM, MIN, MAX, AVG та ін.);
- умови відбору (LIKE, BETWEEN, ALL, ANY та ін.);
- додаткові умови відбору (GROUP BY, HAVING, ORDER BY та ін.).

Як приклад наведені декілька шаблонів SQL-запитів вибірки та додавання даних в якості вихідних даних для кластеризації (див. таблицю).

Приклади шаблонів запитів як вихідні дані для кластеризації

№ запиту	Тип запиту	Клас запиту	Шаблон SQL-запиту
1	SELECT	S_{11}	SELECT c1 FROM t1
2	SELECT	S_{11}	SELECT c1, c2, c3 FROM t1
3	SELECT	S_{12}	SELECT c1, c2 FROM t1 WHERE c1=1 AND c2=2
4	INSERT	I_1	INSERT INTO t1 VALUES ('c1')
5	INSERT	I_1	INSERT INTO t1 VALUES ('c1', 'c2', 'c3')

На виході отримуємо дендрограми, побудовані за двома методами: одиночного зв'язку та повного зв'язку розрахунку дистанції між кластерами (рис. 3).

Дендрограма являє собою дерево, тобто граф без циклів, побудований по матриці мір близькості, і дозволяє зобразити взаємні зв'язки між об'єктами із заданої множини [7]. Метод побудови дендрограми визначає спосіб перерахунку матриці подібності (відмінності) після об'єднання (чи поділу) чергових двох об'єктів у кластер. Гілки дерева — кластери. На кінцях дерева представлені SQL-запити, номери яких вказані в таблиці.



Рис. 3. Дендрограми алгоритмів кластеризації: метод одиночного зв'язку (а); метод повного зв'язку (б)

В обох алгоритмах прості запити типу SELECT увійшли в один кластер, тоді як запит з WHERE опинився на рівень вище. У наведеному прикладі алгоритм повного зв'язку є більш логічним, тому як він об'єднав запити на вибірку (SELECT) в одну гілку, а запити на додавання (INSERT) увійшли до іншої. Відповідно до запропонованих критеріїв можна виділити декілька рівнів класифікації шаблонів запитів.

Первинна класифікація включає 4 типи шаблонів: S_1 — запити типу SELECT, I_1 — запити типу INSERT, U_1 — запити типу UPDATE, D_1 — запити типу DELETE.

Вторинна класифікація дозволяє більш детально розрізнити шаблони запитів, наприклад:

S_{11} — прості запити на вибірку даних з однієї таблиці "SELECT atr1, atr2 FROM tbl1 WHERE atr1=int";

S_{12} — запити на вибірку даних з однієї таблиці з подальшим сортуванням або угрупованням результату "SELECT atr1, atr2, atr3 FROM tbl2 WHERE atr2>int ORDER BY atr3";

S_{13} — запити на вибірку даних з декількох таблиць із зазначенням зв'язків та додаткових умов для результату “SELECT tbl1.atr1, tbl1.atr2, tbl2.atr2, tbl2.atr3 FROM tbl1, tbl2 WHERE tbl1.atr1=tbl2.atr3 and tbl2.atr1=int”.

Результати. Для реалізації запропонованої методики розроблено програмний модуль класифікації шаблонів SQL-запитів “SQL Classification”, який є частиною програмного комплексу класифікації ІС. Множина еталонних шаблонів запитів формується шляхом виділення SQL-запитів з транзакцій TPC за допомогою застосування інструменту тестування продуктивності БД Benchmark Factory for Databases [8]. Під час синтаксичного розбору виконується аналіз структури SQL-запитів, представлених у вигляді дерев реляційної алгебри, метою якого є визначення окремих предикатів запиту. Для проведення аналізу SQL-запиту використовується JSQL Parser [9], який дозволяє перевести SQL-код в ієрархію Java класів (Select, Create, Insert та ін.), а також бібліотека General SQL Parser [10] для побудови XML дерев. Результат розбору SQL-лексем записується до БД під керуванням СКБД з відкритим кодом PostgreSQL. Далі виконується кластеризація отриманої множини шаблонів запитів з метою виявлення класів шаблонів запитів. Результати кластеризації як еталонні шаблони зазначених класів також зберігаються у БД. На рис. 4 представлено архітектуру модуля побудови класів шаблонів запитів.

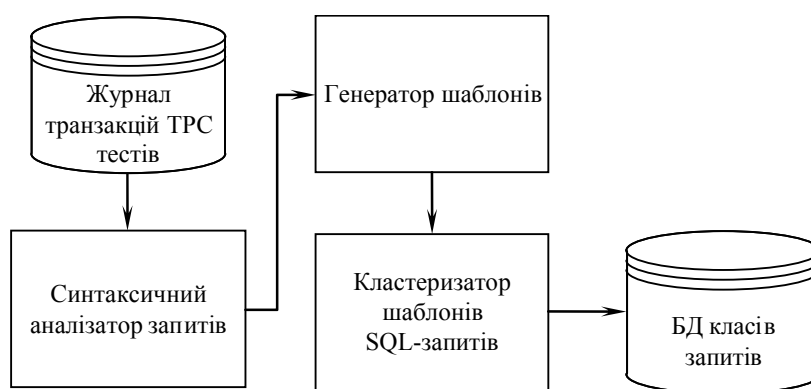


Рис. 4. Модуль побудови класів шаблонів запитів

Використання розробленого комплексу під час підвищення продуктивності інформаційної системи обліку кадрів вищого навчального закладу дозволило розробити тестове середовище, що відповідає БД ІС, яка експлуатується, та виконати тестування з метою пошуку таких значень параметрів налаштування, що дозволили підвищити продуктивність ІС на 41,15 %.

Також отримана множина шаблонів запитів різних класів ІС дозволяє виконувати класифікацію досліджуваної ІС [11] і подальше налаштування СУБД під заданий клас ІС, як системи обробки транзакцій. Метод класифікації ІС включає трьохетапну процедуру виявлення подібності моделей ІС[11]:

- визначення кількості співпадаючих вершин у моделі (збіг шаблонів запитів);
- визначення співпадаючих напрямків переходів в моделі (послідовності виникнення шаблонів запитів) і їх кількості;
- розрахунок пропорційності ймовірностей переходів.

Перший етап методу — розрахунок коефіцієнту співпадаючих шаблонів запитів — передбачає проведення процедури синтаксичного порівняння текстів запитів. Попередня класифікація шаблонів запитів із використанням еталонних класів, представлених в даній статті, дозволяє виключити етап порівняння запитів і тим самим підвищити ефективність запропонованого методу класифікації ІС.

Експериментально доведено, що використання запропонованого методу попередньої класифікації ІС під час її налаштування з метою підвищення продуктивності виявила можливість

виключити етап емпіричного дослідження значень параметрів конфігурації СУБД, як наслідок час настроювання скорочено більш ніж в 10 разів (з 13,8 до 1,36 год).

Висновки. Запропонована методологія побудови шаблонів SQL-запитів для формування бази даних еталонних шаблонів. Як основу для використання методології доцільно використовувати транзакції тестування TPC, що є промисловими стандартами тестування систем обробки транзакцій декількох типів. Алгоритм побудови шаблонів SQL-запитів містить послідовність кроків, починаючи з витягування первинного запиту з транзакції, його синтаксичний та семантичний аналіз, генерація шаблону та його збереження як характерного еталону для певного класу запитів. Для віднесення шаблону до певного типу класу використовується метод ієрархічної кластеризації, який дозволяє поєднати в групи SQL-запити зі схожими властивостями.

Використання шаблонів запитів з моделлю транзакційної поведінки користувача, яка зберігає інформацію про послідовність їх виникнення, дозволяють створювати тести, які більшою мірою відповідають реальним процесам досліджуваної системи, ніж стандартні. Таким чином підвищується достовірність результатів тестування для конкретної ІС, що безпосередньо впливає на ефективність її функціонування, а саме на продуктивність обробки запитів користувачів.

Література

1. Блажко, А.А. Модели для автоматизированной оптимизации производительности систем управления базами данных [Текст] / А.А. Блажко, А.Ю. Левченко, А.С. Пригожев // Радиоэлектронні і комп'ютерні системи. — Харьков, 2010. — № 7(48). — С. 24 — 29.
2. Левченко, А.Ю. Использование скрытых марковских моделей для классификации корпоративных информационных систем [Текст] / А.Ю. Левченко, А.С. Пригожев // Электромашинобудування та електрообладнання. — Одесса, 2010. — № 75. — С. 106 — 112.
3. TPC Benchmark C — Standard Specification, Revision 5.11 [Електронний ресурс] / Transaction Processing Performance Council. — 2010. — 130 p. — Режим доступу: http://www.tpc.org/tpcc/spec/tpcc_current.pdf (Дата звернення: 13.03.2014 p.)
4. TPC Benchmark E — Standard Specification, Revision 1.12.0 [Електронний ресурс] / Transaction Processing Performance Council. — 2010. — 286 p. — Режим доступу: <http://www.tpc.org/tpce/spec/v1.12.0/TPCE-v1.12.0.pdf> (Дата звернення: 13.03.2014 p.)
5. TPC Benchmark H — Standard Specification, Revision 2.14.4 [Електронний ресурс] / Transaction Processing Performance Council. — 2012. — 135 p. — Режим доступу: <http://www.tpc.org/tpch/spec/tpch2.14.4.pdf> (Дата звернення: 13.03.2014 p.)
6. TPC Benchmark W — Standard Specification, Version 1.8 [Електронний ресурс] / Transaction Processing Performance Council. — 2002. — 201 p. — Режим доступу: http://www.tpc.org/tpcw/spec/tpcw_V1.8.pdf (Дата звернення: 13.03.2014 p.)
7. Uprichard, E. Cluster Analysis. Vol. 3: Cluster analysis in practice / E. Uprichard, D.S. Byrne. — London: SAGE, 2012. — 411 p.
8. Benchmark Factory for Databases [Електронний ресурс] / Dell Software. — Режим доступу: <http://www.quest.com/benchmark-factory/> (Дата звернення: 13.03.2014 p.)
9. JSQL Parser [Електронний ресурс] / Режим доступу: <http://jsqlparser.sourceforge.net/home.php> (Дата звернення: 13.03.2014 p.)
10. General SQL Parser [Електронний ресурс] / Gudu Software. — Режим доступу: <http://www.sqlparser.com/> (Дата звернення: 13.03.2014 p.)
11. Левченко, О. Методи моделювання поведінки користувача систем керування базами даних [Текст] / О. Левченко, О. Блажко // Комп'ютерні науки та інженерія: матеріали IV Міжнародної конференції молодих вчених CSE-2010, 25–27 листопада 2010 року, Львів. — Львів: Львівська політехніка, 2010. — С. 72 — 73.

References

1. Blazhko, A.A. Modeli dlya avtomatizirovannoy optimizatsii proizvoditel'nosti sistem upravleniya bazami dannykh [Models for automated optimization of database control systems' performance [Text]] / A.A. Blazhko, A.Yu. Levchenko, A.S. Prigozhev // Radioelektronni i kompiuterni systemy [Radio electronic and computer systems]. — Khar'kov, 2010. — # 7(48). — pp. 24 — 29.

2. Levchenko, A.Yu. Ispol'zovanie skrytykh markovskikh modeley dlya klassifikatsii korporativnykh informatsionnykh sistem [Application of hidden Markov models for classifying corporate information systems [Text]] / A.Yu. Levchenko, A.S. Prigozhev // Elektromashynobuduvannia ta elektroobladnannia [Electrical Machinery and Electrical Equipment]. — Odessa, 2010. — #75. — pp. 106 — 112.
3. TPC Benchmark C — Standard Specification, Revision 5.11 [Electronic resource] / Transaction Processing Performance Council. — 2010. — 130 p. — Available at: http://www.tpc.org/tpcc/spec/tpcc_current.pdf (Access date: 13.03.2014)
4. TPC Benchmark E — Standard Specification, Revision 1.12.0 [Electronic resource] / Transaction Processing Performance Council. — 2010. — 286 p. — Available at: <http://www.tpc.org/tpce/spec/v1.12.0/TPCE-v1.12.0.pdf> (Access date: 13.03.2014)
5. TPC Benchmark H — Standard Specification, Revision 2.14.4 [Electronic resource] / Transaction Processing Performance Council. — 2012. — 135 p. — Available at: <http://www.tpc.org/tpch/spec/tpch2.14.4.pdf> (Access date: 13.03.2014)
6. TPC Benchmark W — Standard Specification, Version 1.8 [Electronic resource] / Transaction Processing Performance Council. — 2002. — 201 p. — Available at: http://www.tpc.org/tpcw/spec/tpcw_V1.8.pdf (Access date: 13.03.2014)
7. Uprichard, E. Cluster Analysis. Vol. 3: Cluster analysis in practice / E. Uprichard, D.S. Byrne. — London : SAGE, 2012. — 411 p.
8. Benchmark Factory for Databases [Electronic resource] / Dell Software. — Available at: <http://www.quest.com/benchmark-factory/> (Access date: 13.03.2014)
9. JSQL Parser [Electronic resource] / Available at: <http://jsqlparser.sourceforge.net/home.php> (Access date: 13.03.2014)
10. General SQL Parser [Electronic resource] / GuduSoftware. — Available at: <http://www.sqlparser.com/> (Access date: 13.03.2014)
11. Levchenko, O. Metody modeliuvannia povedinky korystuvacha system keruvannia bazamy danykh [Information system user's behavior modeling methods] / O. Levchenko, O. Blazhko // Kompiuterni nauky ta inzheneriia: materialy IV Mizhnarod-noi konferentsii molodykh vchenykh CSE-2010, 25-27 lystopada 2010 roku, Lviv [Computer Science and Engineering: Proceedings of the IV International Conference of Young Scientists CSE-2010, November 25-27, 2010, Lviv]. — Lviv, 2010. — pp. 72 — 73.

АНОТАЦІЯ / АННОТАЦИЯ / ABSTRACT

О.Ю. Левченко. Побудова класів SQL-запитів для моделювання поведінки користувачів інформаційних систем і тестування їх продуктивності. Представлено методологію формування бази даних еталонних шаблонів SQL-запитів різних класів, що використовує метод ієрархічної кластеризації і транзакції стандартів ради з засобів обробки транзакцій — Transaction Processing Council (TPC). Еталонні шаблони запитів використовуються для побудови спеціалізованих тестів аналізу продуктивності баз даних інформаційних систем і класифікації поведінки користувачів баз даних інформаційних систем. Розроблено програмний модуль класифікації шаблонів запитів, який було інтегровано в програмний комплекс класифікації та налаштування інформаційних систем. Застосування програмного комплексу дозволило підвищити продуктивність бази даних інформаційної системи із використанням модуля тестування та скоротити час налаштування за рахунок попередньої класифікації.

Ключові слова: реляційні бази даних, шаблони SQL-запитів, тестування баз даних, класифікація інформаційних систем, системи обробки транзакцій.

А.Ю. Левченко. Построение классов SQL-запросов для моделирования поведения пользователей информационных систем и тестирования их производительности. Представлена методология формирования базы данных эталонных шаблонов SQL-запросов разных классов, которая использует метод иерархической кластеризации и транзакции стандартов совета по средствам обработки транзакций — Transaction Processing Council (TPC). Эталонные шаблоны запросов используются для построения специализированных тестов анализа производительности баз данных информационных систем и классификации поведения пользователей баз данных информационных систем. Разработан программный модуль классификации шаблонов запросов, который был интегрирован в программный комплекс классификации и настройки информационных систем. Использование программного комплекса позволило повысить производительность базы данных информационной системы с использованием модуля тестирования и сократить время настройки за счет предварительной классификации.

Ключевые слова: реляционные базы данных, шаблоны SQL-запросов, тестирование баз данных, классификация информационных систем, системы обработки транзакций.

O.Yu. Levchenko. The SQL-queries classes generation for information system user's behavior modeling and systems' performance testing. In article it is represented the methodology for creating the SQL-query templates of different types based on hierarchical clustering and transactions of TPC (Transaction Processing Council) standards. SQL-query tem-

plates can be used to generate the specified tests for database systems performance analyzing and for database system user behavior classification. It is developed the software for query templates classification that was integrated into the software system for information systems classification and configuration. The software applying allowed to increase the database productivity using testing and to reduce the tuning time by pre-classification.

Keywords: relational databases, SQL query templates, database testing, information system classification, transaction processing systems.

Рецензент д-р техн. наук, проф. Одес. нац. політехн. ун-ту Крісілов В.А.

Надійшла до редакції 3 лютого 2014 р.

УДК [004.891.3:62-33]:519.1

О.С. Савельева, д-р техн. наук, доц.,
Д.А. Пурич, інженер,
Одес. нац. політехн. ун-т,
Г.В. Налева, канд. техн. наук, доц., Одес. нац.
мор. акад.

МОДЕЛИРОВАНИЕ СОСТОЯНИЯ РЕЗЕРВИРОВАННЫХ СЕТЕВЫХ СИСТЕМ С ЛАТЕНТНЫМИ ЭЛЕМЕНТАМИ С ПОМОЩЬЮ СКРЫТОЙ МАРКОВСКОЙ МОДЕЛИ

Введение. Как известно, техническая диагностика решает обширный круг задач, в частности, весьма важной является оценка состояния системы, ее надежности. Процедура технической диагностики осуществляется без разборки изделия, а ее основной задачей является распознавание состояния системы в условиях ограниченной информации. Контроль технического состояния системы при эксплуатации, по результатам которого принимается решение о вероятности безотказной работы в заданных условиях с установленными параметрами, является важной составной частью системы обеспечения надежности объекта. При этом решения должны приниматься с учетом соответствующих правил, основываться на методах отказов, а для принятия обоснованных решений привлекаться методы теории статистических решений [1].

Анализ последних исследований и публикаций. Процедуры распознавания в технической диагностике основываются на соответствующих моделях, устанавливающих связь между состояниями технической системы и их отображениями в пространстве диагностических сигналов. В основе многих методов, которые при этом используются, — нахождение математической зависимости между повреждениями системы и ее работоспособностью [1, 2]. Основные существующие подходы (вероятностный и детерминистский) принципиально не отличаются. При детерминистском подходе задача сводится к разделению пространства состояний на области диагнозов; при вероятностном — к построению решающего правила, по которому известные параметры, характеризующие с определенной вероятностью состояние системы, относятся к одному из возможных состояний (признаков). Модели, базирующиеся на статистических характеристиках сигнала, описывают Гауссовы, пуассоновские, марковские, а также по-

DOI: 10.15276/opu.1.43.2014.34

© О.С. Савельева, Д.А. Пурич, Г.В. Налева, 2014