УДК 519.7

М. А. ВОЛК, О. А. КОБЫЛИН, С. Н.САРАНЧА

ФОРМАЛЬНЫЙ АППАРАТ ПОСТРОЕНИЯ ПАРАЛЛЕЛЬНЫХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ОБЪЕКТОВ РАСПРЕДЕЛЕННЫХ БАЗ ДАННЫХ

На современном этапе развития вычислительной техники наблюдается широкое использование распределенных баз данных для построения единых информационных систем. Примерами таких систем являются, например, единая налоговая база данных Украины, базы данных Министерства Внутренних Дел, корпоративные базы данных предприятий и организаций.

К особенностям подобных баз данных можно отнести:

1. Огромный объем хранимой информации.
2. Оригинальность каждой записи (которая соответствует персоне, объекту, и т.п.).
3. Распределение записей по отдельным носителям и подсистемам (например, по территориальному принципу).
4. Большие временные и ресурсные затраты на поиск записей, верификацию при внесении изменений и дополнений.
5. Возможные ошибки операторов при выполнении операций, перечисленных в п.4.

Перечисленные особенности приводят к необходимости создания алгоритмов работы с подобными базами данных, которые учитывают большие ресурсные затраты и неоднозначность при формировании объектов (записей) баз данных.

В качестве примера можно привести базу данных для Отдела виз и регистраций. Объектом базы данных выступает информация об отдельном гражданине, содержащая поля: имя, фамилию, номер и дату выдачи паспорта, возраст, и т.д. При поиске, проверке, замене данных, необходимо идентифицировать текущий объект с объектами всей базы данных. При этом необходимо учитывать, что оператор мог внести ошибку в существующую или в водимую информацию, это может повлечь за собой отрицательный результат по выполнению операций над базой данных, а также к выполнению значительного объема вычислительных ресурсов, затрачиваемых на выполнение указанных действий.

На сегодняшний день существует множество программ и алгоритмов, позволяющих работать с такими базами данных. Однако именно к их использованию сейчас предъявляется много претензий. Одной из причин этого – недостаточная формализация задачи и алгоритмов управления распределенными базами данных. Целью данной статьи является рассмотрение основ формального аппарата постановки и решения задачи распознавания объектов распределенной базы данных.

Дадим формальное описание объектам и условиям задачи.

Совокупность объектов задачи есть кортеж:

$$O=\{S, F, G\},$$

где S – система объектов базы данных: $S = \{S1, S2, \ldots, S_N\}$, N – число элементов базы данных, F – множество функций соответствия двух элементов базы данных $\bar{f}(S_i, S_j)$; G – функция обработки результатов операции соответствия: $G(\bar{f}(S_i, S_j)), i, j = \overline{1, N} \mid i = j$,

Элемент базы данных является совокупностью полей $S_i=\{e_1, e_2, \ldots, e_K\}$, где K – число полей в элементе базы данных. Каждое поле элемента имеет свою длину (размер) $z(e_k)$ и алфавит $A(e_k)$ возможных значений.

Определение 1. Простейшей базой данных называется такая база данных S, число полей элементов которой постоянно: $K(S_i) = K(S_j) \forall i, j \in [1, N]$; размер соответствующих полей одинаков: $z(e_k^{S_i}) = z(e_k^{S_j}), \forall i, j \in [1, N] \& \forall k, l \in [1, K]$; и соответствующие поля определяются одинаковым

алфавитом: $A(e_k) \Leftrightarrow A(e_l) \forall k,l \in [1,K]$.

Понятие одинакового алфавита допускает следующее толкование: одинаковыми алфавитами будем считать любые два алфавита с одинаковыми множествами символов, либо с множествами, с однозначным взаимным соответствием символов (группы символов) этих алфавитов. Такая трактовка позволяет работать с алфавитами разных языков, при условии наличия функции, которая сопоставляет символы алфавита. Примером применения данного положения может служить запись паспортных данных клиента паспортного стола на разных языках (украинском, английском, русском). В этом случае, при наличии правил сопоставления букв (буквосочетаний) алфавитов, можно организовать операции с элементами базы данных, сформированными на разных языках.

Определение 2. Предикативная функция соответствия есть дискретная функция, принимающая значение 1, если все соответствующие поля аргументов равны между собой. В противном случае, она равна 0:

$$\overline{f}(S_i,S_j)=\begin{cases}1, если\ S_i.e_k = S_j.e_k \forall k \in [1,K] \\ 0, если\ S_i.e_k \neq S_j.e_k\ хотя\ бы\ для\ одного\ k \in [1,K]\end{cases} \quad (1)$$

Определение 3. Вероятностная функция соответствия определена в диапазоне [0, 1] и выражает вероятность того, что два элемента базы данных соответствуют одному объекту:

$$\overline{f_p}(S_i,S_j) = P(S_i \Leftrightarrow S_j).$$

Следствие. Исходя из определения вероятностной функции соответствия, ее область допустимых значений $\overline{f_p} \in [0,1]$.

Соответствие элементов базы данных определяется через равенство их полей, следовательно, функция соответствия элементов базы данных, есть композиция функций соответствия полей элементов базы данных. Вероятностная функция соответствия, как правило, используется на промежуточных этапах проведения анализа, а на этапе принятия решения необходим переход от вероятностной функции к предикативной функции соответствия.

Определение 4. Порогом чувствительности функции соответствия называется величина ρ, переводящая значение вероятностной функции соответствия в одно из определенных значений предикативной функции соответствия:

$$\overline{f}(S_i,S_j)=\begin{cases}0, если\ \overline{f_p}(S_i,S_j) < \rho \\ 1, если\ \overline{f_p}(S_i,S_j) \geq \rho\end{cases} \quad (2)$$

Лемма 1. В качестве порога чувствительности функции соответствия элементов базы данных, может выступать произведение вероятностных функций соответствий одноименных полей элементов базы данных:

$$\rho = \overline{f_p}(S_i.e_1,S_j.e_1)\cdot\overline{f_p}(S_i.e_2,S_j.e_2)\cdot\ldots\cdot\overline{f_p}(S_i.e_K,S_j.e_K) \quad (3)$$

Доказательство. Пусть ρ есть величина, определенная выражением (3). На основании следствия из определения 3 областью допустимых значений функции соответствия является интервал [0, 1], тогда, $\overline{f_p}(S_i.e_1,S_j.e_1) \in [0,1]$, $\overline{f_p}(S_i.e_2,S_j.e_2) \in [0,1]$, ..., $\overline{f_p}(S_i.e_K,S_j.e_K) \in [0,1]$. Следовательно, произведение функций: $\overline{f_p}(S_i.e_1,S_j.e_1)\cdot\overline{f_p}(S_i.e_2,S_j.e_2)\cdot\ldots\cdot\overline{f_p}(S_i.e_K,S_j.e_K) \in [0,1]$, то есть ρ в этом случае: $\rho \in [0,1]$. Вероятностная функция соответствия $\overline{f_p}(S_i,S_j) \in [0,1]$, следовательно,

значение величины ρ попадает в область допустимых значений функции $\overline{f_p}(S_i, S_j)$, а значит, может быть порогом согласно определению 4.

Определение 5. Два элемента базы данных имеют сходство, если их предикативная функция соответствия равна 1.

Определение 6. Функция обработки результатов операции соответствия G возвращает значение целевой функции оценки выполнения одной их команд (или ее части) управления базой данных.

Например, команда поиска аналогичного объекта базы данных, в процессе ее выполнения, может возвратить несколько объектов базы данных, имеющих сходство (определяемое функциями соответствия), среди которых надо выделить одну (несколько) для возврата пользователю, этим реализуется цель выполнения команды.

Используя введенные определения можно выделить те объекты задачи распознавания, которые допускают параллельное выполнение (рис. 1).

Распределенная база данных подразумевает расположение подмножества ее элементов (S^N) на удаленных вычислительных ресурсах. Таким образом, появляется возможность организации парал-

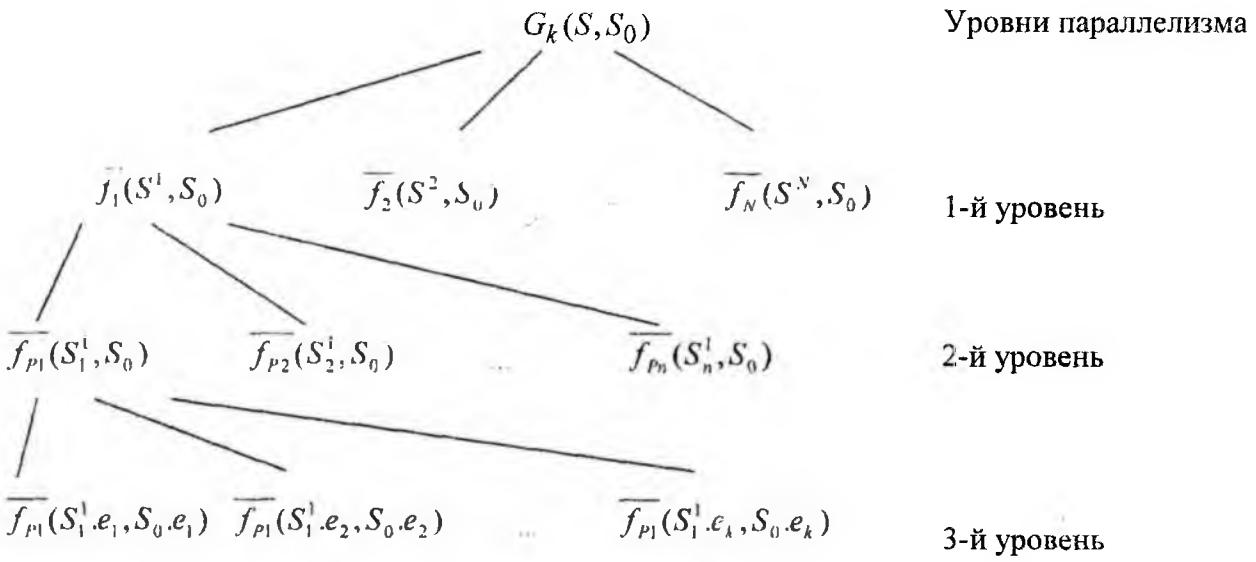

Рис. 1

лелизма первого уровня как совокупности подзадач распознавания сегментов распределенной базы данных. Параллелизм второго уровня основан на том факте, что результаты выполнения функций соответствия пары выбранных элементов базы данных не зависят друг от друга. И, наконец, третий уровень параллелизма определяется независимостью, в общем случае, выполнения функций соответствия полей элементов.

Введенные в статье определения позволяют представить процесс распознавания в параллельном виде, что приводит к возможности построения формального аппарата параллельных алгоритмов распознавания объектов распределенных баз данных. Решение данной задачи предполагается раскрыть в дальнейших работах.

Список литературы: 1. *Клини С.* Математическая логика. М.: Мир, 1973. 480 с. 2. *Горбачев В.А., Волк М.А., Бабаев А.П.* Методы декомпозиции моделей непрерывных систем для моделирования в условиях распределенных ресурсов // Радиотехника и информатика. 1998. №1. С.35-38. 3. *Колесников Д. О., Пославский С.А, Шабанов-Кушнаренко Ю. П.* Идентификация начальных логических понятий // Проблемы бионики. 2000. Вып.52 С.9-18.

Поступила в редколлегию 1.07.2001