Дрождин В.В., Баканов А.Б. Грамматика описания домена фамилий. – Вопросы радиоэлектроники. Серия «Электронная вычислительная техника». – 2007, Вып.1. – С. 77-82.

В.В. Дрождин
А.Б. Баканов

УДК 681.3

ГРАММАТИКА ОПИСАНИЯ ДОМЕНА ФАМИЛИЙ

Рассматривается формализация представления данных в домене фамилий на основе формальной грамматики. Предложен критерий и способ его вычисления, позволяющий сравнивать языки и грамматики.

Использование семантических свойств в представлении и организации данных в базе данных (БД) позволяет существенно повысить эффективность функционирования систем баз данных и автоматизированных информационных систем (АИС), построенных на их основе. Легко установить, что основные семантические свойства классов объектов отражаются в структуре синтаксического представления информации об этих объектах. Поэтому серьезным препятствием для обеспечения высокой эффективности функционирования систем БД являются:

- использование в них универсальных языков представления различных данных на внутреннем (физическом) уровне;
- использование универсальных (единых для всей системы и для различных данных) алгоритмов обработки данных.

Например, в реляционной модели данных, получившей в настоящее время наибольшее распространение, данные рассматриваются как атомарные (целостные, неделимые) объекты, вследствие чего объем данных в БД возрастает прямо пропорционально объему накапливаемой информации о предметной области. В сетевой и иерархической моделях данных объем одних и тех же данных в БД несколько меньше, чем в реляционной, но громоздкость обрабатываемых структур и отсутствие мощного языка запросов теоретико-множественного типа сводят на нет это преимущество.

Доменно-ориентированный подход позволяет снять эти проблемы, рассматривая домены как отдельные самостоятельные системы БД, взаимодействующие между собой на основе единого стандартизованного интерфейса. Обособленность и специализация доменов позволяют в полном объеме реализовать принцип физической независимости данных в БД.

Домен содержит информацию о множестве объектов одного типа и, следовательно, все данные в домене имеют одинаковую или незначительно различающуюся структуру. Это позволяет разрабатывать внутреннюю организацию домена специально ориентированную на хранение и обработку данных именно с требуемой структурой, что повышает эффективность функционирования домена на несколько порядков.

Для выявления структуры данных элементов домена может использоваться аналитический подход, разработанный в математической

лингвистике для анализа естественных и формальных языков. При этом любые системы данных считаются открытыми как вверх (создание суперсистем, в которых данные системы являются элементами), так и вниз (любой элемент является сложным, состоящим из элементов более низкого уровня, вплоть до отдельных символов). Поэтому на основе анализа элементов домена можно выявлять их более или менее устойчивую структуру и формировать существенно более эффективную внутреннюю организацию домена.

Рассмотрим внутреннюю организацию домена фамилий, в основе которой будет находиться формальная грамматика.

1. Неформальное описание домена фамилий

Домен фамилий представляет собой некоторое множество слов-фамилий. Для конкретности будем рассматривать русские фамилии, представляемые в виде:

$$z_1z_2\ldots z_n \quad \text{или} \quad z_1z_2\ldots z_n{-}z_{n+1}z_{n+2}\ldots z_m{-}z_{m+1}z_{m+2}\ldots \tag{1}$$

где $z_1, z_{n+1}, z_{m+1}, \ldots$ – прописная буква кириллицы;

$z_2, \ldots, z_n, z_{n+2}, \ldots, z_m, z_{m+2}, \ldots$ – строчная буква кириллицы.

Домен может содержать также иностранные фамилии, записанные на русском языке в виде (1) или в виде:

$$z_1z_2\ldots z_n{\text{␣}}z_{n+1}z_{n+2}\ldots z_m{\text{␣}}z_{m+1}z_{m+2}\ldots \tag{2}$$

где ‘␣’ – символ пробела.

Каждой фамилии сопоставим бинарный признак-пол со значениями ‘ж’ – женский и ‘м’ – мужской.

Так как между фамилией и полом задано простое отношение ассоциации (соответствия), то домен фамилий относится к простым доменам.

Таким образом, полное значение элемента домена фамилий представляется в виде:

$$z_1z_2\ldots z_n\delta\# \quad \text{или} \quad z_1z_2\ldots z_n\gamma z_{n+1}z_{n+2}\ldots z_m\gamma z_{m+1}z_{m+2}\ldots\delta\#, \tag{3}$$

где γ – символ ‘–’ или ‘␣’;

δ – признак-пол со значениями ‘ж’ или ‘м’;

– финальный (завершающий) символ элемента домена.

2. Теоретико-множественная модель домена фамилий

Для записи домена фамилий используется конечный **алфавит** символов:

$$A_ф = A_1 \cup A_2 \cup A_3 \cup \{\#\}, \tag{4}$$

где A_1 = {‘А’…‘Я’} – алфавит прописных букв кириллицы;

A_2 = {‘а’…‘я’} – алфавит срочных букв кириллицы;

A_3 = {‘-’, ‘␣’} – алфавит специальных символов;

‘#’ – символ, являющийся признаком завершения фамилии.

На $A_ф$ задается **свободная полугруппа** $\mathcal{F}^*$ [1] в виде множества всех конечных последовательностей символов (слов), на котором определена ассоциативная и некоммутативная бинарная операция конкатенации

(сцепления). Из $\mathcal{F}^*$ выделим такое подмножество $\mathcal{F}$, в котором каждое слово $x_ф \in \mathcal{F}$ представляется в формате:

$$x_ф = z_1z_2\ldots z_nδ\#\quad \text{или}\quad x_ф = z_1z_2\ldots z_nγz_{n+1}z_{n+2}\ldots z_mγz_{m+1}z_{m+2}\ldots δ\#,$$

где $z_1, z_{n+1}, z_{m+1}, \ldots \in A_1$;

$z_2, \ldots, z_n, z_{n+2}, \ldots, z_m, z_{m+2}, \ldots \in A_2$;

$γ \in A_3$;

$δ \in \{‘ж’, ‘м’\} \subset A_2$.

Тогда **язык** $L_ф$, представляющий конкретный домен фамилий, является подмножеством:

$$L_ф \subseteq \mathcal{F}. \qquad (5)$$

Множество $\mathcal{F}$ есть универсальный язык представления всех возможных фамилий, а полугруппа $\mathcal{F}^*$ − универсальный язык над $A_ф$.

Учитывая форматы слов $x_ф \in L_ф$, в $\mathcal{F}^*$ можно выделить свободную подполугруппу $\mathcal{F}_2^* \subset \mathcal{F}^*$ со свободным порождающим множеством A_2.

Для количественной оценки полугруппы $\mathcal{F}^*$ или любого ее подмножества введем оценочную функции φ. Например, $φ(\mathcal{F}^*)$ – количество всех символов в полугруппе $\mathcal{F}^*$, $φ(L_ф)$ – количество всех символов в языке $L_ф$, $φ(x_ф)$ – количество символов в слове $x_ф$.

Таким образом, домен фамилий представляется формальным языком $L_ф = \{x_{фi} \mid 0 \le i \le |L_ф|\}$, находящемся в отношении $L_ф \subseteq \mathcal{F} \subset \mathcal{F}^*$. Здесь $|L_ф|$ - количество слов-фамилий в языке $L_ф$.

3. Контекстная грамматика, порождающая домен фамилий

При представлении домена фамилий с помощью некоторой порождающей грамматики будем использовать **алфавит символов**:

$$\widetilde{A}_ф = T_ф \cup N_ф,$$

где $T_ф = A_ф$ – конечное фиксированное множество терминальных символов;

$N_ф$ – множество нетерминальных символов.

Правила подстановки

$$P_ф = \{p_ф\}$$

являются множеством продукций вида $p_ф : x \to y$, каждая из которых на основе слова x порождает слово y.

Начальный символ грамматики $S_ф$ является последовательностью слов вида $S_ф = \{λ_i \mid 0 \le i \le n\}$,

где $λ_i$ – слово, позволяющее восстановить одно или несколько слов $x_ф \in L_ф$;

n – количество слов в последовательности $S_ф$.

Грамматика $G_ф$, порождающая язык $L(G_ф)$, задается в виде [2]:

$$G_ф = (S_ф, T_ф, N_ф, P_ф), \qquad (6)$$

Если в объявлении $S_ф$ параметр n = 0, то $S_ф = \varnothing$ и грамматика $G_ф$ порождает пустой язык $L(G_ф) = \varnothing$.

Для построения грамматики $G_ф$ будем использовать аналитический подход [3], заключающийся в анализе языка $L_ф$ с целью выявления структуры слов, состава их элементов и отношений между этими элементами.

В начальном состоянии будем считать, что для языка $L_ф$ определено **единичное разбиение** E, в каждую клетку e_i которого помещается одно и только одно слово $x_{фi}$, т.е. $x_{фi} \in e_i$. При этом $E = \{e_i\}$, $e_i = \{x_{фi}\}$, $|E| = |L_ф|$, $L_ф = \bigcup_{i=0}^{|L_ф|} e_i$, $0 \le i \le |L_ф|$. Язык $L_ф$ становится языком с парадигматической структурой ($L_ф$, E), но без морфологии, так как слова языка являются целостными объектами, не имеющими внутренней структуры.

Представлению языка ($L_ф$, E) соответствует грамматика $G_{0ф} = (S_{0ф}, T_ф, N_{0ф}, P_{0ф})$, в которой $S_{0ф} = E$, $\lambda_i = e_i = x_{фi}$. $P_{0ф}$ содержит только правила описания формата представления слов $x_{фi}$:

$p_1 : x_{фi} \to z\delta\#, \ 1 \le i \le |L_ф|$

$p_2 : z \to z_1z_2z_3\ldots z_n \mid z\gamma z_1z_2z_3\ldots z_n$

$p_3 : z_1 \to А \mid Б \mid \ldots \mid Я$

$p_4 : z_2z_3\ldots z_n \to а \mid б \mid \ldots \mid я$

$p_5 : \gamma \to - \mid \text{␣}$

$p_6 : \delta \to ж \mid м$

$N_{0ф}$ включает нетерминальные символы из правил подстановки $p_1 \ldots p_6$.

Грамматика $G_{0ф}$ будет порождать язык $L(G_{0ф}) = L_ф$ и иметь размерность $|L(G_{0ф})| = |L_ф|$ и $\varphi(L(G_{0ф})) = \varphi(L_ф)$.

Учитывая, что функция φ дает количественную оценку последовательности символов, то ее можно применять как к отдельным правилам $p_ф$, так и ко всему множеству правил $P_ф$. Поэтому становится возможным количественное сравнение языков с порождающими их грамматиками, а также сравнение грамматик между собой. Например, количественное сравнение языка $L_ф$ и порождающей грамматики $G_{0ф}$ дает соотношение $\varphi(G_{0ф}) = \varphi(L_ф)$. Это означает, что грамматика $G_{0ф}$ представляет собой запись слов языка $L_ф$ в определенной последовательности.

Совершенствование грамматики $G_{0ф}$ может осуществляться двумя путями:

- построением более крупных слов на основе обобщения слов $x_ф \in L_ф$;
- уменьшением размера слов $x_ф \in L_ф$ на основе декомпозиции на составные части и классификации этих составных частей.

Оценку степени совершенства грамматики будем производить на основе количества символов, необходимых для записи грамматики. Будем считать грамматику G'' более совершенной относительно грамматики G', если $\varphi(G'') < \varphi(G')$ при $L(G'') = L(G')$. Совершенство грамматики G'' относительно грамматики G' тем выше, чем больше разность $\Delta n = \varphi(G') - \varphi(G'')$.

Оценивая возможные пути совершенствования грамматики $G_{0ф}$ по критерию уменьшения размера Δn можно ожидать получение наиболее совершенной грамматики $G_{1ф}$ при использовании обобщения слов, т.к. в этом случае строится более крупное разбиение E_1, в каждую клетку $e_{1,j}$, которого помещаются слова из одной или нескольких клеток $e_i \in E$, $e_{1,j} = \{x_{фi} \mid x_{фi} \in e_i, p_k : \alpha_j \to x_{фi}\}$.

Библиография

1. Скорняков Л.А. Элементы алгебры: Уч. пос. для вузов. – М.: Наука, 1986. – 240 с.

2. Рейуорд-Смит В.Дж. Теория формальных языков. Вводный курс. – М.: Радио и связь, 1988. – 128 с.

3. Маркус С. Теоретико-множественные модели языков. - М.: Наука, 1970. – 332 с.