Дрождин В.В., Тобольченко В.М. Отношение семантической эквивалентности форматов представления данных. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей IX Междунар. научно-техн. конф. – Пенза: ПДЗ, 2009. – С. 27-33.

ОТНОШЕНИЕ СЕМАНТИЧЕСКОЙ ЭКВИВАЛЕНТНОСТИ ФОРМАТОВ ПРЕДСТАВЛЕНИЯ ДАННЫХ

В.В. Дрождин, В.М. Тобольченко

Пензенский государственный педагогический университет
им. В.Г. Белинского,
г. Пенза, Россия

Для повышения разрешающей способности АИС и выполнения семантических преобразований данных на синтаксическом уровне предложено использовать формат данных. Приведены определения, свойства и примеры формата данных и поля форматов, подробно рассмотрено отношение семантической эквивалентности форматов данных.

Drozhdin V.V., Tobolchenko V.M. The relation of the semantic equivalence of data representation format.

To increase the resolving capacity of AIS and performing the semantic data transformations on the syntactic level the use of data format is suggested. The definitions, properties and examples of data format and field of formats are given. The relation of the semantic equivalence of data representation format is viewed in detail.

Огромные достижения в области компьютерной техники и технологий программирования в последние годы почти не коснулись взаимодействия пользователей с автоматизированными информационными системами (АИС). Представление и обработка информации в АИС осуществляется на основе типов данных, определяющих синтаксическое представление и допустимые операции над ними. Компьютер до сих пор принципиально не может отличить, например, дату рождения от фамилии человека или адреса проживания, если все они заданы строкой символов, а механизм распознавания не запрограммирован. При этом у пользователей нет никаких средств, чтобы научить этому компьютер, а у компьютера нет средств научиться этому самостоятельно.

Поэтому необходимо разработать средства и методы представления и обработки данных, повышающие разрешающую способность АИС и упрощающие процесс взаимодействия с ней пользователей и других систем. Для этого будем использовать формат данных.

Формат данных – это форма синтаксического представления данных, отражающая основные закономерности их построения, определяемые семантикой и использованием. Например, представление данных в краткой или полной форме, в стандартизованной форме или в форме, требуемой определенным документом или пользователем. Таким образом, формат отражает структуру и согласованную форму представления данных, достаточную для их использования при решении практических задач.

Конструктивно формат данных представляется обобщенной грамматикой, выделяющей в значении данных определенные компоненты (поля) и задающей их

значения в обобщенном виде. Например, в значении данных «адрес» могут быть выделены поля: почтовый индекс, область, район, населенный пункт, улица, дом, корпус и квартира. При этом почтовый индекс задается числовым значением, область, район и улица – символьными значениями с признаками «обл.», «р-н» и «ул.», населенный пункт – символьным значением с признаками «гор.», «пос.» и др., корпус – числовым или символьным значением с признаком «кор.», а дом и квартира – числовым значением с признаком «д.» и «кв.» соответственно.

Для обработки различных данных АИС использует набор форматов, который целесообразно организовать в виде системы форматов с определенной структурой, позволяющей эффективно представлять форматы для хранения и обработки.

Система форматов – множество форматов с определенными отношениями между ними. Между форматами могут существовать следующие отношения: включение, наследование, семантическая эквивалентность и др.

Подробнее рассмотрим отношение семантической эквивалентности форматов.

Форматы F_1 и F_2 назовём ***семантически эквивалентными***, если они представляют данные эквивалентные на семантическом уровне, но различающиеся синтаксически, и будем обозначать это как

$$F_1 \overset{c}{\equiv} F_2. \qquad (1)$$

Формально отношение эквивалентности форматов (1) определим следующим образом: пусть $F_1(d_1)$ и $F_2(d_2)$ синтаксические представления значений данных d_1 и d_2 соответственно. Тогда форматы F_1 и F_2 будут эквивалентны тогда и только тогда, когда будут семантически эквивалентны значения данных d_1 и d_2, т.е.

$$F_1 \overset{c}{\equiv} F_2 \Leftrightarrow d_1 \equiv d_2.$$

Например, представление числа 3,123456 с точностью до трех знаков:

$$F_1 \overset{c}{\equiv} F_2 \Leftrightarrow 3{,}123456 \equiv 3{,}123.$$

Частным случаем эквивалентности значений данных, но и более сильной ее формой, является равенство данных, которое также соответствует отношению эквивалентности форматов данных.

Отношение семантической эквивалентности форматов обладает следующими свойствами [2]:

1) рефлексивность: $F_1 \overset{c}{\equiv} F_1$;

2) симметричность: если $F_1 \overset{c}{\equiv} F_2$, то $F_2 \overset{c}{\equiv} F_1$;

3) транзитивность: если $F_1 \overset{c}{\equiv} F_2$ и $F_2 \overset{c}{\equiv} F_3$, то $F_1 \overset{c}{\equiv} F_3$.

Например, дата 14 октября 2009 г. может быть задана в двух форматах: «14.10.2009» и «10/14/2009». Обе записи представляют одну и ту же дату (семантика), но имеют различную синтаксическую структуру. В данном случае говорится о семантической эквивалентности двух форматов даты.

Так как формальной составляющей формата является грамматика [1], то к формату можно применить операции преобразования грамматик. Операции преобразования применимы к грамматике $G = (T, N, P, S)$, если множество правил P приведено к следующему виду:

$$
\begin{aligned}
&P = \{S \to \nu_1 \nu_2 ... \nu_n \\
&p_1 : \nu_1 \to ... \\
&p_2 : \nu_2 \to ... \\
&... \\
&p_n : \nu_n \to ... \\
&\}
\end{aligned}
\qquad (2)
$$

Принципиальной особенностью грамматик вида (2) является то, что правая часть начального символа *S* состоит только из нетерминальных символов. Каждое правило имеет свой идентификатор и обладает семантическими (синтаксическими) свойствами. Представление грамматик в виде (2) отражает структуру порождаемых цепочек символов.

Преобразование грамматики *G* вида (2) в грамматику *G'* такого же вида будем задавать формулой

$G \xrightarrow{\theta} G'$ или $G' = \theta(G)$,

где θ – допустимая операция над грамматикой *G*.

Приведем основные операции преобразования грамматик.

1. Замена цепочки символов – это операция получения грамматики *G'* из грамматики *G* вида (2) путем замены правила p_k грамматики *G*, порождающего цепочку символов α, на правило p_k' грамматики *G'*, порождающего цепочку символов β:

$$G' = \theta_{зам.}(G, p_k, p_k'),$$

где $p_k : \nu_k \to \alpha$;

$p_k' : \nu_k \to \beta$.

Например, два формата представления адреса эквиваленты при замене цепочки символов «ул.» на значение «улица»:

[улица][*название улицы*][ ][*№ дома*][-][*№ квартиры*] $\overset{c}{\equiv}$
[ул.][*название улицы*][ ][*№ дома*][-][*№ квартиры*],

так как «ул. Мира 12-34» ≡ «улица Мира 12-34».

2. Удаление цепочки символов является частным случаем операции замены цепочки символов, так как в качестве β выступает пустая цепочка символов ε:

$$G' = \theta_{удал.}(G, p_k) = \theta_{зам.}(G, p_k, p_k'),$$

где $p_k : \nu_k \to \alpha$;

$p_k' : \nu_k \to \varepsilon$.

Например, два формата представления адреса эквиваленты при удалении цепочек символов «д.» и «кв.»:

[ул.][*название улицы*][ ][д.][*№ дома*][-][кв.][*№ квартиры*] $\overset{c}{\equiv}$
[ул.][*название улицы*][ ][*№ дома*][-][*№ квартиры*],

так как «ул. Мира д. 12 – кв. 34» ≡ «ул. Мира 12-34».

3. Вставка цепочки символов позволяет получить грамматику *G'* путём добавления цепочки символов α в позицию *k* начального символа *S* грамматики *G*:

$$G' = \theta_{встав}(G, \nu', k).$$

Например, два формата представления адреса эквиваленты при вставке цепочки символов «улица»:

[*название улицы*][ ][*№ дома*][–][*№ квартиры*] $\overset{c}{\equiv}$
[улица][*название улицы*][ ][*№ дома*][–][*№ квартиры*],
так как «Мира 12-34» ≡ «улица Мира 12-34».

4. Модификация числовых цепочек позволяет получить грамматику G' из грамматики G вида (2) путем изменения цепочки символов x, являющейся числовым значением и порождаемой нетерминальным символом ν_k правила p_k, на цепочку символов y, порождаемую из x по закону f:

$$G' = \theta_{ч\_мод}(G, p_k, p_k', f),$$

где $p_k : \nu_k \to \alpha,\ \nu_k.val = x;$
$p_k' : \nu_k \to \beta,\ \nu_k.val = y;$
$y = f(x).$

Например, два формата представления единиц измерения веса эквиваленты при замене единиц измерения с « кг» на « гр» и модификации значения веса x:

[x][кг] $\overset{c}{\equiv}$ [*x*1000*][гр],
так как «2 кг» ≡ «2000 гр».

5. Изменение порядка следования цепочек позволяет получить грамматику G' из грамматики G вида (2) путем взаимного изменения позиций двух нетерминальных символов в начальном символе S грамматики G:

$$G' = \theta_{поряд.}(G, \nu_k, \nu_\ell).$$

Например, два формата представления даты эквиваленты при изменении порядка следования цепочек «*дд*» и «*мм*»:

[*дд*][/][*мм*][/][*гг*] $\overset{c}{\equiv}$ [*мм*][/][*дд*][/][*гг*],
так как «14/02/09» ≡ «02/14/09».

При помощи приведенных операций можно производить эквивалентные преобразования грамматик (форматов).

Форматы позволяют, например, представлять единицы измерения в различных системах единиц (СИ, СГС и др.), время – 24- и 12-часовой формах исчисления и т.д. Так, одна и та же величина длины для европейского пользователя будет представлена в метрах (сантиметрах), а для американского – в дюймах. Система может отслеживать предпочтения пользователей в выборе форматов в процессе ввода данных.

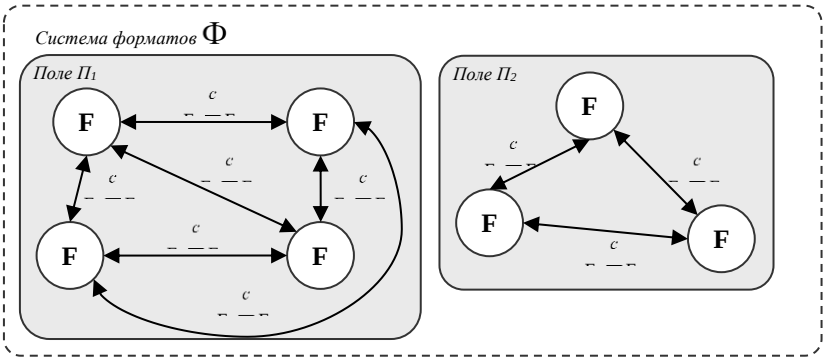
Отношение семантической эквивалентности форматов позволяет АИС реализовать следующие функции:

преобразовывать данные к согласованному виду;

выбирать оптимальный формат для хранения данных;

представлять данные пользователям в требуемой форме.

Совокупность семантически эквивалентных форматов назовём **полем форматов** (рисунок).



Система форматов, включающая два поля

На рисунке схематично представлена система форматов Φ, состоящая из 7 форматов, разделенных на два поля:

$\Pi_1 = \{F_1, F_2, F_3, F_4\}$;

$\Pi_2 = \{F_5, F_6, F_7\}$.

В системе форматов поля играют роль категорий, разделяющих форматы по семантическому признаку. Так, например, в информационной системе форматы, представляющие единицы измерения, могут быть разделены по физическим величинам [3]:

поле единиц измерения массы;

поле единиц измерения расстояния;

поле единиц измерения температуры и т.д.

Воспользуемся правилами преобразования форматов и построим поле эквивалентных форматов единиц измерения расстояний (таблица).

Поле единиц измерения расстояния

НАЗВАНИЕ	ОБОЗНАЧЕНИЕ	ГРАММАТИКА
МЕТР	$F_М$	$G_м$: $S \to \nu_1\nu_2\nu_3$ $p_1 : \nu_1 \to x$ $p_2 : \nu_2 \to \_$ $p_3 : \nu_3 \to м$, ГДЕ $x \in R, x \geq 0$.
САНТИМЕТР	$F_{СМ}$	$G_{см} = [\theta_{ч\_мод}(G_м, p_1, p_1', (\nu_1.val/100)) \circ$ $\circ \theta_{зам}(G_м, p_3, p_3')]$, ГДЕ $p_3' \to см$
КИЛОМЕТР	$F_{КМ}$	$G_{км} = [\theta_{ч\_мод}(G_м, p_1, p_1', (\nu_1.val * 1000)) \circ$ $\circ \theta_{зам}(G_м, p_3, p_3')]$, ГДЕ $p_3' \to км$
ДЮЙМ	$F_{дюйм}$	$G_{дюйм} = [\theta_{ч\_мод}(G_м, p_1, p_1', (\nu_1.val * 39,37)) \circ$ $\circ \theta_{зам}(G_м, p_3, p_3')]$, ГДЕ $p_3' \to дюйм$

Выделим основные методы построения (выявления) отношения семантической эквивалентности в системе форматов:

1) инициализация отношений на этапе создания системы – ввод первоначальных знаний начиная от самые простых, например систем единиц измерения, и заканчивая более сложными (адреса, документы и т.д.);

2) обучение системы – метод, предполагающий наличие учителя, обучающего систему. В качестве учителя может выступать пользователь. Например, пользователь может явно задать отношение эквивалентности между двумя форматами «1.01.02 = 1/01/02»;

3) самоорганизация – метод, при котором система без специфического воздействия извне формирует структуру форматов [4 – 6], а также отношения между ними, в том числе отношение эквивалентности.

Таким образом, даны определения ключевым понятиям: формат, система форматов, поле форматов; перечислены отношения, которые могут быть установлены между форматами; подробно рассмотрено отношение эквивалентности, его структура, свойства и функции.

Библиографический список

1. Карпов Ю.Г. Теория и технология программирования. Основы построения трансляторов. – СПб. : БХВ-Петербург, 2005. – 272 с.

2. Кострикин А.И. Введение в алгебру. – М. : Наука, 1977.

3. Сена Л.А. Единицы физических величин и их размерности. – М., 1965. – 304 с.

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

5. Дрождин В.В. Метаграмматика описания форматов данных в информационных доменно-ориентированных средах // Проблемы информатики в образовании, управлении, экономике и технике : сб. ст. IV Всерос. науч.-техн. конф. – Пенза, 2004. – С. 43–45.

6. Тобольченко В.М. Построение шаблонов для множества значений переменной длины // Материалы 55-й науч. студ. конф. – Пенза, 2006. – С. 82 – 83.