Дрождин В.В., Тобольченко В.М. Анализ различий синтаксического представления наборов однородных данных. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей VIII Всерос. научно-техн. конф. – Пенза: ПДЗ, 2008. – С. 14-18.

АНАЛИЗ РАЗЛИЧИЙ СИНТАКСИЧЕСКОГО ПРЕДСТАВЛЕНИЯ НАБОРОВ ОДНОРОДНЫХ ДАННЫХ

В.В. Дрождин, В.М. Тобольченко

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

В современных автоматизированных информационных системах (АИС) представление и обработка информации осуществляется на основе типов данных, определяющих синтаксическое представление и допустимые операции над ними. При этом компьютер принципиально не может отличить, например, дату рождения от фамилии человека или адреса проживания, если все они заданы строкой символов. Поэтому для распознавания семантики значения данных ее необходимо указывать компьютеру явно, например, *фамилия = «Иванов»,* или задать значение *«Иванов»* в поле ФАМИЛИЯ. Однако это имеет два негативных последствия:

1) пользователям АИС приходится приспосабливаться к возможностям системы вместо того, чтобы система приспосабливалась к потребностям пользователей;

2) вследствие низкой разрешающей способности системы и достаточно невысокой надежности человека разработка и использование АИС является очень трудоемким процессом.

Поэтому целесообразно разработать некоторый способ представления и обработки данных, повышающий разрешающую способность АИС и упрощающий процесс взаимодействия с ней пользователей и других систем.

Исходя из всего разнообразия информации, определим и рассмотрим символьные представления набора данных, отражающих множество семантически однородных объектов предметной области. В качестве примеров приведем общеизвестные наборы данных: номера телефонов, серии и номера паспортов, адреса, даты, международный стандартный номер книги (ISBN), регистрационные знаки транспортных средств, множество натуральных чисел, множество рациональных чисел и т.д.

Целью данного исследования является изучение символьных представлений различных широко используемых наборов данных, а также их свойств и закономерностей, таких, как алфавит, длина слова, формат, эквивалентность форм представления данных, семантика, распознавание.

Алфавитом называется конечное непустое множество A. Его элементы называются символами (буквами) [1]. Для каждого типа данных определен свой алфавит.

Словом (цепочкой, строкой, string) в алфавите A называется конечная последовательность элементов из A.

Слово состоит из полей и разделителей. Разделитель – это последовательность или отдельный символ, определяющий структуру (формат) набора данных. Такие

символы не меняются в последовательностях и часто сохраняют свои позиции. Обозначим алфавит символов разделителей как A^P . Все последовательности символов от начала до первого разделителя, между разделителями и от последнего разделителя до конца строки являются полями. Обозначим алфавит символов, записываемых в полях как $A^П$. Тогда алфавит терминальных символов будет иметь вид:

$A = A^P \cup A^П$.

Пусть $M_Д$ – множество значений, представляющих даты, например, 13.02.08, 2.12.06, 28.09.98 и т.д. Тогда

$A_Д = \{0,1,2,3,45,6,7,8,9,.\}$

$A_Д^П = \{0,1,2,3,45,6,7,8,9\}$

$A_Д^P = \{.\}$.

Длина слова x, обозначаемая как $|x|$, есть число символов в x, причем каждый символ считается столько раз, сколько раз он встречается в x [2].

Длина слов может быть фиксированной (постоянной) или переменной. Например, список почтовых индексов содержит строки фиксированной длины (6 символов), а список фамилий – переменной длины.

При рассмотрении множеств последовательностей переменной длины необходимо указать две характеристики размера значений данных: минимальную и максимальную длины цепочек.

Выявление синтаксической структуры (формата) представления набора данных фиксированной длины проще, чем данных переменной длины. Поэтому хранение, поиск и обработка данных, имеющих фиксированную длину, осуществляются более простыми методами, чем данных переменной длины.

Формат данных – это синтаксические правила (грамматика) построения данных. Формат представления набора данных, обладающих высокой степенью подобия, можно задать относительно небольшой совокупностью грамматических правил. Набор данных сложной структуры может быть представлен системой форматов с иерархической или сетевой структурой. Для выявления формата данных может использоваться аналитический подход, разработанный в математической лингвистике для анализа естественных и формальных языков [2].

Пусть значения набора данных $x \in X$ представляют собой конечные последовательности символов $x = a_1 a_2 .. a_n$, где $a_i \in A$. Все множество значений данных X составляет язык $L_X = \{x_i \mid i \in I\}$, где I – номерное множество.

Основываясь на анализе языка L_X, АИС должна разработать эффективную грамматику $G_x = (S,T,N,P)$, удовлетворяющую условию $L_X = L(G_X)$, где $T = A$ – множество терминальных символов, N – множество нетерминальных символов, $P = \{p\}$ – множество правил подстановки, заданных в виде продукций $p : x \to y$, порождающих по слову x слово y , $S = \{s_i \mid i \in I\}$ – начальный символ грамматики, являющийся последовательностью слов, представляющих слова из L_X в G_X [3]. Для выполнения условия $L_X = L(G_X)$ необходимо, чтобы каждое слово $x_i \in L_X$ однозначно порождалось по слову $s_i \in S$ и множеству правил P .

Для обеспечения высокой эффективности алгоритмов порождения грамматики G_X и обработки данных на ее основе целесообразно строить грамматику G_X в форме контекстно-свободной грамматики (КС-грамматики).

Однако для набора сложных значений данных, компоненты (поля) которых будут незначительно отличаться синтаксически, построить КС-грамматику, по-видимому, не удастся.

Применение форматов, более точно определяющих синтаксис данных, предоставляет следующие дополнительные возможности:

1) более точное графическое представление данных при вводе;

2) эквивалентное представление данных в разных формах (например, разные единицы измерения).

Эквивалентность форм представления данных. Одно и то же значение может быть представлено в различных формах. Например, даты могут быть представлены в формах: «дд.мм.гг», «дд.мм.гггг», «гг.мм.дд», «гггг.мм.дд», «мм.дд.гг», «мм.дд.гггг», «дд-месяц-гг», «дд-месяц-гггг».

Кроме этого, использование свойства эквивалентности данных, представленных в разных формах, позволяет задавать отношения между форматами представления одних и тех же данных, например, веса: «X т» = «10*X ц» = «1000*X кг». Из нескольких эквивалентных форм представления данных для внутренней организации данных система может выбрать форму, наиболее эффективную для обработки и требующую минимального объема памяти. Следовательно, возможность эквивалентного представления данных в разных формах позволяет с помощью форматов определить универсальный способ преобразования данных.

Разнообразие значений. При рассмотрении проблемы построения формата представления множества значений данных, целесообразно различать множество возможных значений данных $\overline{X}$ и набор значений данных X, отражающий реально существующие объекты. Чем меньше множества значений данных X и $\overline{X}$, тем потенциально более эффективные грамматики можно разработать для их представления. Например, логический тип данных определяет всего два значения: «истина» и «ложь», поэтому целесообразно преобразовывать такие значения в 0 и 1. При $|X| \approx |\overline{X}|$ надежность грамматического представления X будет достаточно высокой, так как вероятность появления значений из $(\overline{X} - X)$ может быть достаточно низкой. В случае $|X| << |\overline{X}|$ для представления X может быть разработан очень эффективный формат представления данных. Однако при добавлении в X новых значений из $(\overline{X} - X)$ часто будет возникать необходимость в изменении формата представления данных.

Семантика. Рассмотрим свойство отражения семантики данных в их синтаксическом представлении. Как отмечалось выше, слова разбиваются на поля и разделители, которые следуют в определенном порядке. Семантика однородных объектов фиксируется в структуре данных, отражающих эти объекты. Очень часто структура данных соответствует только конкретному набору семантически однородных объектов. Поэтому выявление синтаксической структуры данных имеет определенную семантическую ценность. В противном случае семантика

данных должна быть задана явно (например, *фамилия = «Иванов»*), порядком следования элементов данных или каким-либо другим способом.

Пользователь при вводе данных будет задавать правильные данные, если не будет нарушать структуру формата. В случае совершения ошибки система может отреагировать на это по одному из двух сценариев: 1) запросить подтверждение правильности ввода, 2) самостоятельно заполнить поля в правильном порядке.

Распознавание. Ведение алфавитов, форматов (грамматик), деление структуры слов на поля и разделители, учет длины слов позволяют системе в большом количестве случаев практически безошибочно распознавать данные и распределять их по соответствующим наборам данных.

Однако могут возникать ситуации множественной интерпретации значения данных на основе нескольких форматов. Это соответствует КС-грамматике представления набора данных сложной структуры.

Таким образом, на основе синтаксического анализа имеющихся данных АИС может самостоятельно формировать форматы представления данных и на их основе осуществлять более корректное отражение предметной области и взаимодействие с пользователями.

Библиографический список

1. Пентус, А.Е., Пентус, М.Р. Математическая теория формальных языков. – http://www.intuit.ru/department/algorithms/mathformlang/1/

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

3. Линьков, В.М., Дрождин, В.В., Лушникова, Е.В. Порождение грамматики описания данных в информационных доменно-ориентированных средах // Проблемы информатизации в образовании, управлении, экономике и технике: сборник статей III Всероссийской науч.-техн. конференции. – Пенза, 2003. – С. 8-11.