РАЗРАБОТКА МЕТОДА ОПИСАНИЯ СЕМАНТИКИ АТРИБУТОВ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Комар Ф. В.

Липецкий Государственный Технический Университет

Липецк, Россия

В задачах интегрирования баз данных часто возникает проблема оценки сходства объектов [1]. В большинстве случаев такого рода оценка сходства может базироваться на некоторых семантических характеристиках объектов [2]. Так, например, наиболее примитивной семантической характеристикой атрибутов отношений можно считать тип атрибута. Однако при интегрировании комплексных баз данных, такой характеристики недостаточно. Возникает проблема разработки более сложных семантических характеристик атрибутов, на базе которых в дальнейшем можно разрабатывать меры сходства объектов баз данных. В данной работе будет предложена семантическая характеристика атрибутов отношений на базе строковых шаблонов.

Шаблон – общеизвестный образец, трафарет. Шаблоны используются для сжатого описания некоторого множества объектов, без необходимости перечисления всех экземпляров этого множества.

Пусть дано множество объектов (экземпляров) некоторого типа. Пусть на этом множестве заданы правила определения шаблонов и язык шаблонов L - это формальный язык определения шаблонов. Каждый шаблон $j \in L$ определяет набор экземпляров j_e, которые удовлетворяют данному шаблону. Множество j_e является подмножеством множества всех возможных экземпляров U [3].

Опишем синтаксис и структуру шаблонов, которые будем использовать для описания строковых данных. Разобьем символы в иерархически упорядоченные группы (см. рис. 1).

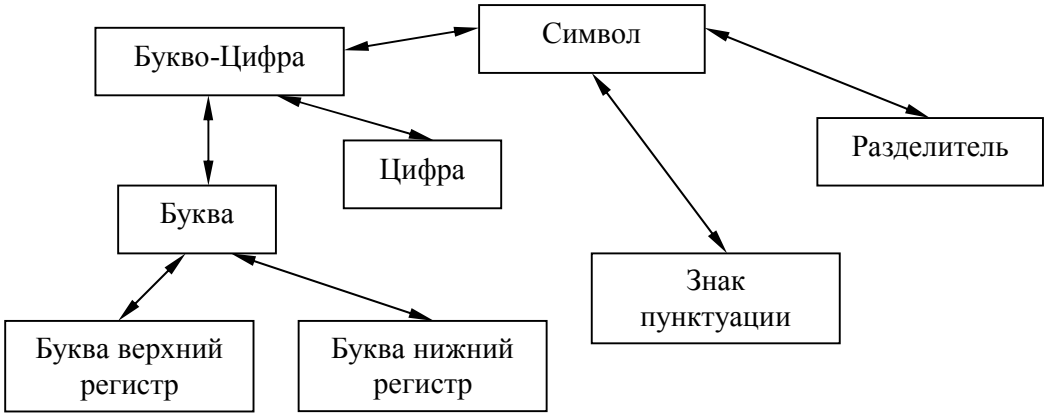

Рис. 1. Иерархия символов строкового шаблона.

В квадратных скобках будем обозначать группы символов, которые могут присутствовать на текущей позиции строки. Например [а, б, в] – множество букв а, б, в. Конструкция вида [а, б, в]{n, m} – означает, что символы а, б, в встречаются в количество от n до m. Конструкция вида [а, б, в]{n, } – означает, что символы а, б, в встречаются в количество не менее n. Конструкция вида [а, б, в]{ , m} – означает, что символы а, б, в встречаются в количество не более m. Отметим, что в квадратных скобках может так же присутствовать некоторый шаблон, который в данном случае будем называть подшаблоном.

Для удобства использования и в соответствии с рисунком 1 введем следующие обозначения групп символов:

\p{Lower} – множество букв нижнего регистра: [а, б, в, …, я];

\p{Upper} – множество букв верхнего регистра: [А, Б, В, …, Я];

\p{ASCII} – множество любых символов;

\p{Alpha} – множество букв [а, б, в, … , я, А, Б, В, … , Я];

\p{Digit} – множество цифр [0, 1, 2, … , 9];

\p{Alnum} – множество букв и цифр;

\p{Punct} – знак пунктуации [!, ", #, $, %, &, ', (,), *, +, ,, -, ., /, :, ;, <, =, >, ?, @, [, \,], ^, _, `, {, |, }, ~];

\p{Space} – множество разделителей [\t, \n, \f, \r, \s];

Как было показано выше, любой шаблон определяет некоторое множество строк. И можно считать, что данный шаблон является некоторым семантическим описанием этого множества строк. Очевидно, что один шаблон не может полностью описать все семантические особенности данного множества строк, однако некоторую семантическую

значимость шаблон, безусловно, несет. С одной стороны шаблон тем лучше описывает множество строк, чем больше строк из этого множества удовлетворяют шаблону. С другой стороны шаблон тем лучше описывает множество строк, чем больше строк, не принадлежащих данному множеству, не удовлетворяют этому шаблону. Семантической значимостью можно считать некоторую обобщенную численную оценку, удовлетворяющую указанным выше свойствам. Можно так же предположить, что при определенных условиях некоторое множество шаблонов в совокупности будет иметь семантическую значимость для множества строк.

Для примера рассмотрим множество строк вида: Имя Фамилия. Естественным образом можно сказать, что шаблон вида $\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}\p\{Space\}\{1,\}\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}$ имеет некоторую семантическую значимость. Очевидно так же, что указанный выше шаблон не представляет полностью семантику множества строк указанных выше. Более того, для предложенного примера можно составить целое множество шаблонов, которые будут с тем или иным уровнем семантической значимости описывать множество указанных строк. Например:

$\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}\p\{Space\}\{1,\}\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}$

$\p\{Alpha\}\{1, \}\p\{Space\}\{1, \}\p\{Alpha\}\{1, \}$

$\p\{Alpha\}\{1, \}\p\{Space\}\{1, \}$

$\p\{Alpha\}\{1, \}[-]\{ ,1\}\p\{Alpha\}\{1, \}$

и т.д.

Очевидно, что для множества строк, можно отыскать такой шаблон, которому будут удовлетворять все строки данного множества, однако при этом семантической значимости у этого шаблона будет не велика. Так например семантическая значимость шаблона вида $\p\{ASCII\}\{1, \}$ будет гораздо меньше чем семантическая значимость шаблона вида $\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}\p\{Space\}\{1,\}\p\{Upper\}\{1,1\}\p\{Lower\}\{1,\}$.

Любая реляционная база данных содержит некоторое множество атрибутов, а так же множество конкретных значений каждого атрибута [4]. Пусть $A=\{a_1,a_2,a_3,...,a_n\}$ - множество всех атрибутов базы данных. Пусть D_i - множество значений атрибута a_i, $\overline{D}=\{D_1,D_2,D_3,...,D_n,\}$ - набор, множеств значений атрибутов, j - некоторый шаблон. Рассмотрим функцию:

$$pF(j, D_i) = \frac{freq(j, D_i)}{|D_i|} \quad (1)$$

Где $freq(j, D_i)$ - определенная выше функция, которая возвращает количество строк из множества D_i, которые удовлетворяют шаблону j, а $|D_i|$ - объем множества D_i.

Функция $pF(j, D_i) = \frac{freq(j, D_i)}{|D_i|}$ дает численную оценку того, насколько точно шаблон описывает строки, которые принадлежат рассматриваемому домену. Значения функции лежат на отрезке [0, 1]. В дальнейшем эту величину будем кратко называть частотой появления шаблона j на множестве D_i.

Определим функцию:

$$pAF(j, D') = \frac{1}{m}\sum_{j=1}^{m} pF(j, D_j) \quad (2)$$

Где $D' = \{D_1, D_2, D_3, ..., D_m\}$ - набор множеств значений атрибутов. Указанная функция дает усредненное значение численной оценки того, насколько точно шаблон описывает строки, принадлежащие соответствующим множествам строк.

Определим функцию:

$$pV(j, D_i, D_i') = \max(pF(j, D_i) - pAF(j, D_i'), 0) \quad (3)$$

Где D_i - множество значений i-ого атрибута, $D_i' = \{D_1, D_2, D_3, ..., D_{i-1,} D_{i+1,} ..., D_n\}$ - набор всех множеств значений атрибутов, кроме i-ого. Значение функции тем выше, чем больше экземпляров множества i-ого атрибута удовлетворяют шаблону j и чем меньше среднее значение количества экземпляров остальных атрибутов удовлетворяющих шаблону. Значения функции лежат на отрезке [0, 1]. Максимальное значение функция принимает в том случае, когда все значения i-ого атрибута удовлетворяют шаблону j, и ни один экземпляр остальных атрибутов не удовлетворяет шаблону j.

Примем значение функции pV как численное выражение семантической значимости атрибута A_i относительно атрибутов $A_i' = \{A_1, A_2, A_3, ..., A_{i-1,} A_{i+1,} ..., A_n\}$ в контексте шаблона j.

Для множества шаблонов $U = \{j_1, j_2, j_3, ..., j_n\}$ определим функцию семантической значимости, как среднее значение семантической значимости каждого шаблона в отдельности:

$$psV(U, D_i, D_i') = \frac{1}{n} \sum_{j=1}^{n} pV(j_j, D_i, D_i') \quad (4)$$

Таким образом, множество шаблонов может считаться некоторой семантической характеристикой атрибута реляционной базы данных. Для построения такого множества необходимо решить задачу максимизации функции семантической значимости. Разработка метода решения такого рода задачи позволит автоматически строить семантическую характеристику атрибутов реляционных баз данных.

СПИСОК ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

1. W. Hasselbring. Information system integration. //Communications of the ACM, 43(6)33-38, 2000.
2. Цаленко М. Ш. Моделирование семантики в базах данных. - М.: Наука, 1989. - 287 с.
3. Фридл Дж. Регулярные выражения, 2-е издание. – Спб.: Питер, 2003. – 464 с.
4. Дейт К. Дж. Введение в системы баз данных, 7-е издание. - Пер. с англ. - М.: Издательский дом Вильямс, 2001. - 1072 с.