АЛГОРИТМ МАКСИМИЗАЦИИ ФУНКЦИИ ОЦЕНКИ СЕМАНТИЧЕСКОЙ ЗНАЧИМОСТИ СТРОКОВОГО ШАБЛОНА

Комар Ф. В.

Липецкий Государственный Технический Университет

Липецк, Россия

В настоящее время активно ведется работа по созданию методов автоматизированного интегрирования баз данных. [1] В большинстве случаев эти методы базируются на оценках семантического сходства объектов. Однако описание семантики объектов является нетривиальной задачей, которая до сих пор окончательно не решена. Таким образом, исследования методов описания семантики объектов являются актуальной задачей. [2]

Рассмотрим возможность использования строковых шаблонов в качестве семантической характеристики множества семантически сходных строк. В качестве языка строковых шаблонов будем использоваться общеизвестный язык регулярных выражений. [3] Любой строковый шаблон определяет некоторое множество строк. И можно считать, что строковый шаблон является некоторым семантическим описанием множества строк. Семантической значимостью можно считать некоторую обобщенную численную оценку, характеризующую то, насколько данный шаблон точно описывает заданное множество строк.

Пусть D_i - множество строк, обладающих сходной семантикой, $\overline{D} = \{D_1, D_2, D_3, ..., D_n, \}$ - некоторый набор, множеств строк, j - некоторый шаблон. Определим функцию:

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

где $freq(j, D_i)$ - функция, которая возвращает количество строк из множества D_i, которые удовлетворяют шаблону j, а $|D_i|$ - объем множества D_i. Значение функции pF 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 строковых значений.

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. Примем значение функции pV как численное выражение семантической значимости шаблона j относительно множества строк D_i в контексте набора множеств строк $D_i' = \{D_1, D_2, D_3, ..., D_{i-1,} D_{i+1,} ..., D_n\}$.

Таким образом, задача семантической характеристики некоторого множества строк относительно набора множеств других строк может быть сведена к задаче максимизации функции семантической значимости шаблона.

Для решения задачи максимизации функции семантической значимости используем генетический алгоритм [4] представленный на рисунке 1.

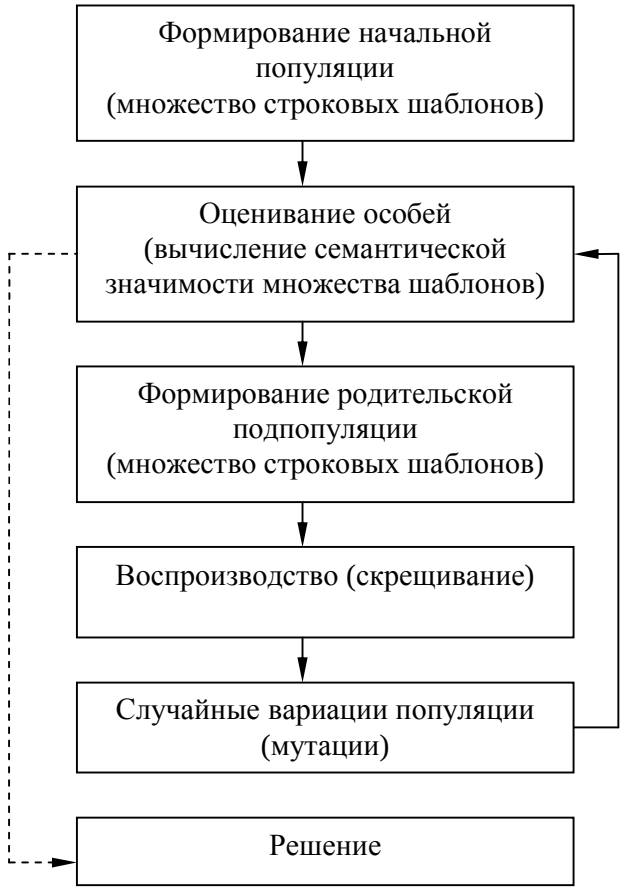


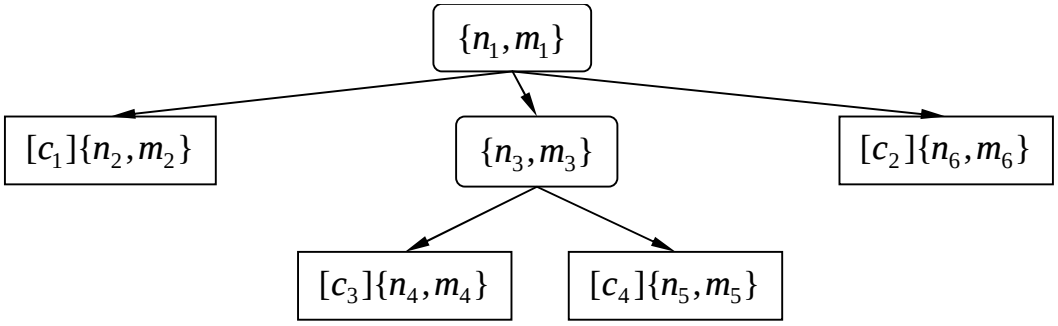
Рис. 1. Генетический алгоритм максимизации функции семантической значимости.

Использование генетического алгоритма подразумевает представление в генетическом виде информации о шаблоне, поэтому прежде чем перейти к описанию разработанного алгоритма, определим способ кодирования информации о шаблоне в виде генов.

В общем случае структура шаблона, может содержать любое количество подшаблонов, а сам шаблон может быть представлен в виде дерева. Узлами дерева будут являться

подшаблоны, которые в свою очередь содержат другие подшаблоны. Листья дерева будут представлять собой шаблоны, которые не содеражат подшаблоны, но характеризуются множеством допустимых символов. Пример древовидной структуры шаблона показан на рисунке 2.

Пример древовидной структуры шаблона



Строковое представление шаблона

$$([c_1]\{n_2, m_2\}([c_3]\{n_4, m_4\}[c_4]\{n_5, m_5\})\{n_3, m_3\}[c_2]\{n_6, m_6\})\{n_1, m_1\}$$

Рис. 2. Древовидная структура шаблона.

В терминах эволюционных алгоритмов каждый подшаблон представляет собой хромосому. Множество генов объединенных в древовидную структуру будут представлять шаблон, а в терминах эволюционного поиска – особь. Хромосома может состоять из различного количества генов. При этом гены определяют множество допустимых символов подшаблона в случае, если данный подшаблон является листом в дереве подшаблонов, или определяют набор подшаблонов в случае, если данный подшаблон содержит другие подшаблоны. Значение минимального и максимального количества вхождений данного подшаблона так же кодируются в виде генов.

Так как основная задача поиска – отыскание шаблонов, наиболее точно описывающих определенный атрибут в контексте множества других атрибутов, то естественным образом можно определить фитнес функцию как функцию оценки семантической значимости атрибута в контексте множества атрибутов.

Начальную популяцию будем формировать на основе множества значений рассматриваемого атрибута. Каждое значение атрибута может быть закодировано в виде шаблона следующим образом:

На основании каждого символа значения атрибута формируется шаблон. Каждый подшаблон шаблона представляет собой лист в дереве подшаблонов, множество символов

представлено одним текущим символом значения атрибута, максимальное и минимальное количество вхождений подшаблонов равно единице.

Определим оператор скрещивания как случайный обмен хромосомами между двумя особями. В терминах шаблонов, такого рода обмен будет представлять собой обмен некоторыми подшаблонами между двумя деревьями подшаблонов.

Рассмотрим следующие операции над шаблонами:

Добавление подшаблона – операция, добавляющая в дерево подшаблонов новый подшаблон.

Удаление подшаблона – операция удаляющая из дерева подшаблонов подшлаблон.

Изменения минимального количества вхождения подшаблона – изменение параметра подшаблона, характеризующего минимальное вхождения подшаблона.

Изменения максимального количества вхождения подшаблона – изменение параметра подшаблона, характеризующего максимальное вхождения подшаблона.

Уточнение множества символов подшаблона – замена текущего множества символов подшаблона на множество символов, стоящих ниже в иерархии групп символов.

Обобщение множества символов подшаблона – замена текущего множества символов подшаблона на множество символов, стоящих выше в иерархии групп символов.

Добавление символа в множество символов подшаблона – добавление символа, стоящего на том же уровне иерархии символов, что и остальные допустимые символы подшаблона.

Удаление символа из множества символов подшаблона – удаление символа из множества допустимых символов подшаблона.

Определим оператор мутации как случайное применение одной из вышеописанных операций к случайной хромосоме особи. В простейшем случае будем полагать применения любой операции равновероятным.

Предложенный выше алгоритм позволяет отыскать строковый шаблон, который в контексте рассматриваемых множеств строк дает максимум значения функции семантической значимости.

Таким образом, предложен метод описания семантики множества строк с помощью строковых шаблонов, определена функция численной оценки семантической значимости шаблона, а так же предложен алгоритм максимизации данной функции. Строковые шаблоны, которые дают максимум функции семантической значимости, могут быть рассмотрены как семантическая характеристика множества строк.

Список литературных источников

1. Глеб Лодыженский. Шлюзы как средство интеграции баз данных. // Открытые системы, №2, 1999.
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. Курейчик, В.М. Генетические алгоритмы / Л.А. Гладков, В.М. Курейчик, В.В. Курейчик. – М.: Физматлит, 2006.