УДК 004.891:510.62

ЗАДАЧА ПРИВЕДЕНИЯ ПОРЕКВИЗИТНОГО АДРЕСА К СТРОКОВОЙ ФОРМЕ

Гладков Сергей Львович[1],
e-mail: gladkovs@list.ru,
[1]Общество с ограниченной ответственностью «Айгео», г. Красноярск, Россия

В статье исследуется тема формального описания и нормализации общепринятого понятия «адрес» в базе данных. Актуальность работы обусловлена высокой степенью изменчивости адресов и необходимостью создания значительного числа грамматических правил для их формального описания и нормализации. В статье предлагается подход обобщения правил этого описания. Сформулирован критерий эквивалентности множеств порождаемых адресных строк. Разработана система правил формальной грамматики, которая преобразует набор значений реквизитов адреса в строковую форму. Адресная строка рассматривается как лингвистическая переменная, источником для которой является пореквизитный адрес, преобразованный по правилам порождающей грамматики. В свою очередь, каждый пореквизитный адрес рассматривается как результат преобразования соответствующей адресной строки при помощи правил распознающей грамматики. С применением метода логической категоризации процесса преобразования конкретного адреса построено дерево адресов, пути которого определяют структуру каждой адресной строки. Предложена теоретическая модель автоматического преобразования пореквизитного нормализованного адреса в строковую форму. Полученные результаты могут быть полезны проектировщикам информационных систем, содержащих данные адресного описания или взаимодействующих с ними.

Ключевые слова: пореквизитный адрес, строковый адрес, модель адреса, порождающая грамматика, критерий эквивалентности множеств адресов

THE ISSUE OF CONVERTING THE REQUIRED ADDRESS TO A STRING FORM

Gladkov S.L.[1],
e-mail: gladkovs@list.ru,
[1]LLC "AYGEO", Krasnoyarsk, Russia

The article explores the topic of formal description and normalization of the generally accepted concept of "address" in the database. The relevance of the work is due to the high degree of variability of addresses and the need to create a significant number of grammatical rules for their formal description and normalization. The article proposes an approach to summarize the rules of this description. The criterion of sets equivalence of generated address strings is formulated. A system of formal grammar rules has been developed that converts a set of address details values into a string form. The address bar is considered as a linguistic variable, the source of which is the required address, transformed according to the rules of the generative grammar. In turn, each request address is considered as the result of converting the corresponding address bar using the rules of the recognizing grammar. Using the method of logical categorization of the conversion process of a specific address, an address tree has been built, the paths of which determine the structure of each address bar. A theoretical model is proposed for the automatic conversion of a mandatory normalized address into a string form. The results obtained can be useful for designers of information systems containing address description data or interacting with them.

Keywords: a mandatory address, a string address, an address model, a generative grammar, an equivalence criterion for sets of addresses

DOI 10.21777/2500-2112-2024-2-65-81

Введение

В настоящее время существует большое число информационных систем, которые включают в себя данные об адресах объектов недвижимости, физических и юридических лиц в форме набора значений реквизитов. Причем набор реквизитов и их значений в разных информационных системах часто не совпадает. Такой подход не соответствует высокой степени изменчивости адресов, обусловленным, например, укрупнением населенных пунктов, а также ростом числа садовых и огороднических некоммерческих товариществ (СНТ и ОНТ) и др. В отличие от пореквизитного представления адреса в строковом представлении изменяются «на лету». С другой стороны, адрес в форме списка реквизитов удобно использовать как ключ для организации взаимодействия различных информационных ресурсов. Адрес в строковой форме таким свойством не обладает. Поэтому, а также вследствие других причин, адреса существуют в двух формах: пореквизитной и строковой. В результате возникла острая необходимость в механизме автоматического преобразования адреса из одной формы в другую при соблюдении эквивалентности преобразованного адреса исходному.

Целью данной статьи является построение теоретической модели автоматического преобразования пореквизитного нормализованного адреса в строковую форму.

Основным методом исследования является логическая категоризация процесса преобразования конкретного адреса. Для преобразования адресов также используются методы порождающей и распознающей грамматик.

1. Введение в формальный анализ адресных строк

В статье продолжается развитие темы формального описания и нормализации общепринятого понятия «адрес» (адрес здания, адрес регистрации гражданина и т.п.). В предыдущих статьях [1; 2] адрес рассматривался как последовательность значений реквизитов, т.е. $a = \{r_1, r_2, \ldots r_n\}$, где n – количество реквизитов, множество всех адресов A представляло собой подмножество декартова произведения множеств реквизитов адресов, т.е.

$$A \subset R_1 \times R_2 \times \ldots \times R_n,$$

где R_i – множество значений i-го реквизита, $1 \le i \le n$.

В настоящей работе будут рассматриваться адреса a^s в строковом формате, множество которых будут обозначаться $A^s \mid a^s \in A^s$, вопросы взаимосвязи A^s и A и, в частности, построения отношения эквивалентности между ними будет обозначаться $A^s \sim A$. При этом каждую адресную строку a^s множества A^s будем рассматривать как лингвистическую переменную [3], источником для которой является пореквизитный адрес a, преобразованный по правилам порождающей грамматики AG. В свою очередь, каждый пореквизитный адрес a рассматривается как результат преобразования соответствующей адресной строки a^s при помощи правил некой распознающей грамматики A^sG. Построенная таким образом эквивалентность множеств A^s и A позволяет анализировать адресные строки в форме управленческого списка [4, с. 67], а также в более узкой форме протоклассификатора адресов, для которого порожденные реквизиты представляют собой *идентифицирующие характеристики* [5, с. 56].

2. Отношения эквивалентности на адресных списках

Прежде чем продолжать говорить об эквивалентности между множествами строковых и пореквизитных адресов $A^s \sim A$, необходимо ввести критерий, с помощью которого эквивалентность будет устанавливаться. Такой критерий будет строиться из предположения о представимости множеств адресов в виде управленческих списков, эквивалентность которых рассмотрена в статье «Формальные

свойства совместимости списков» [4]. Будем считать, что множества адресов эквивалентны, если эквивалентны соответствующие им управляющие списки.

Списки, построенные на основе множеств адресов, будем называть адресными списками. Поэтому наша задача сведется к рассмотрению вопросов эквивалентности двух и более адресных списков.

Каждый адресный список может быть построен как на полном множестве адресов, так и на любой его части. Поэтому построенный критерий будет применим к доказательству эквивалентности подмножеств адресов и их реквизитов.

2.1. Отношения эквивалентности на множестве адресов

Начнём с рассмотрения отношений эквивалентности, заданных на всём множестве адресов A, представленном в форме адресного списка.

Лемма об эквивалентности ключевых реквизитов. На каждом адресном списке может быть задано не меньше, чем $n\text{-}k+1$ отношений эквивалентности, где n – совокупное число реквизитов контекста и ключа адресного списка, k – число реквизитов ключа списка.

В дальнейшем рассуждении учитывается наличие порядка на множестве реквизитов адреса [1, с. 59].

Действительно, на списке адресов, как на управленческом списке, так и на любом множестве, естественным образом определяется полное отношение эквивалентности $\sim$ по признаку принадлежности к нему. То есть любые два элемента такого списка считаются эквивалентными [6, с. 52]. При этом, предикат принадлежности адресному списку имеет вполне определенную форму [4, с. 63]:

$$\boldsymbol{P}(\boldsymbol{A}) \equiv (\boldsymbol{R}_1 = r_1) \wedge (\boldsymbol{R}_2 = r_2) \wedge \ldots \wedge (\boldsymbol{R}_k = r_k),$$

где $\{\boldsymbol{R}_i \mid 1 \le \boldsymbol{i} \le \boldsymbol{k}\}$ [1] – подмножество адресных реквизитов из состава идентифицирующих характеристик, а $(\boldsymbol{R}_i = r_i)$ – одноместный предикат [7, с. 92] над i-м реквизитом списка адреса.

Можно сказать, что $\boldsymbol{P}(\boldsymbol{A})$ – общий предикат, описывающий критерий принадлежности произвольного адреса a к A, а значит, задает отношение эквивалентности. Предикат элемента списка представляет собой объединение общего и ключевого предикатов:

$$\boldsymbol{P}(\boldsymbol{a}) \equiv (\boldsymbol{R}_1 = r_1) \wedge (\boldsymbol{R}_2 = r_2) \wedge \ldots \wedge (\boldsymbol{R}_k = r_k) \wedge (\boldsymbol{R}_{k+1} = r_{k+1}) \wedge \ldots \wedge (\boldsymbol{R}_n = r_n)$$

$$\text{или } \boldsymbol{P}(\boldsymbol{a}) \equiv \boldsymbol{P}(\boldsymbol{A}) \wedge (\boldsymbol{R}_{k+1} = r_{k+1}) \wedge \ldots \wedge (\boldsymbol{R}_n = r_n),$$

где $\{R_i \mid 1 \le i \le n\}$ – полный набор адресных реквизитов, из которых k имеют общие значения для всех элементов списка, а $R_{k+j} \mid k < j \le n-k$ – это ключевые реквизиты элементов адресного списка A, каждый из которых определён на всем множестве записей адресного списка.

Определим отношение эквивалентности на множестве значений R_{k+1} по признаку принадлежности к нему.

Далее рассмотрим множество значений первого ключевого реквизита $R_{k+1} = \{r_{k+1,1}, r_{k+1,2}, \ldots, r_{k+1,l(k+1)}\}$, где $l(k+1)$ – количество значений этого реквизита. Каждое значение этого множества разбивает множество значений реквизита R_{k+2} на непересекающиеся подмножества $\{R_{k+2,1}, R_{k+2,2}, \ldots, R_{k+2,l(k+1)}\}$ и, тем самым, задает отношение эквивалентности как на множестве значений R_{k+2}, так и на всём адресном списке. При этом соответствующий предикат будет иметь вид $(R_{k+1} = r_{k+1}) \Rightarrow R_{k+2}$.

Аналогично, множество значений ключевого реквизита $R_{k+j-1} = \{r_{k+1,1}, r_{k+1,2}, \ldots, r_{k+1,l(k+j-1)}\}$ задаёт разбиение следующего реквизита $\boldsymbol{R}_{k+j}$ на $l(k+j-1)$ подмножеств – $\{R_{k+j,1}, R_{k+j,2}, \ldots, R_{k+j,l(k+j-1)}\}$.

[1] Символом R_i в зависимости от контекста обозначается как название i-го реквизита, так и множество его значений. Так, Ri в одноместном предикате (Ri = ri) – это название реквизита, а в выражении ri ∈ Ri – множество значений реквизитов. В тех случаях, когда смысл символа окажется непонятен без уточнения имя, реквизита будет обозначаться R_i^N, а множество его значений R_i^D.

В этом случае отношение эквивалентности будет определяться предикатом, соединяющим атомарные высказывания для ключевых реквизитов, предшествующих R_{k+j} при помощи операции конъюнкции

$$\vee_1^{j-1}\left(R_{k+v}=r_{k+v}\right)\Rightarrow R_{k+j}.$$

И, наконец, отношение эквивалентности на множестве значений реквизита R_n может быть задано следующим предикатом:

$$\wedge_1^{n-k-1}\left(R_{k+j}=r_{k+j}\right)\Rightarrow R_n.$$

Итак, на адресном списке было построено *n-k+1* отношений эквивалентности, включающих одно полное отношение и *n-k* отношений на взаимозависимых множествах значений ключевых реквизитов.

Следствие. Отношение эквивалентности на множестве значений ключевого реквизита задает также отношение эквивалентности на всём множестве записей адресного списка.

Для доказательства достаточно показать, что отношение эквивалентности, заданное на множестве значений ключевого реквизита, разбивает всё множество записей адресного списка на непересекающиеся подмножества.

Отношение эквивалентности на множестве значений ключевого реквизита R_1 определяется по признаку принадлежности к нему, а это значит, что разбиение на этом множестве состоит из самого этого множества. Поэтому его расширением является полное отношение эквивалентности на всём адресном списке, которое строится по признаку принадлежности к этому списку.

Отношение эквивалентности на множестве значений ключевого реквизита $\boldsymbol{R_j}$ разбивает его на непересекающиеся подмножества $\left\{R_{j,1},R_{j,2},\ldots,R_{j,l(j-1)}\right\}$. Разобьём множество записей адресного списка на подмножества $\left\{A_1,A_2,\ldots,A_{l(j-1)}\right\}$ так, чтобы значения ключей каждой записи в A_v *j*-го реквизита содержали значения из подмножества $R_{j,v}$, значения меньших реквизитов содержали соответствующие значения из предиката для $R_{j,v}$. Допустим теперь, что $A_v\cap A_u\neq\varnothing$, тогда записи из этого пересечения должны иметь одинаковые значения в ключевых реквизитах с номерами от 1 до *j*. Но по определению разбиения реквизита R_j все значения, обладающие таким свойством, принадлежат только одному подмножеству разбиения. Это значит, что подмножество A_v или подмножество A_u не соответствует исходным подмножествам разбиения значений ключевого реквизита R_j.

Здесь следует сделать несколько замечаний. Во-первых, набор реквизитов адресного списка не ограничен лишь ключевыми реквизитами. Во-вторых, адресный список, при постоянном контексте, может содержать более чем один набор ключевых реквизитов. При этом только один из таких наборов представляет актуальный ключ, а остальные, если они есть, считаются кандидатами в состав ключа адресного списка.

2.2. Отношения эквивалентности между адресными реквизитами

Общность атомарных высказываний в контекстах значений двух и более реквизитов позволяют говорить о существовании между ними отношений эквивалентности. Часть из них являются следствием отношений эквивалентности на множестве адресов. Но есть и особые случаи отношений эквивалентности, связывающих только подмножества значений реквизитов.

Определение префикса предиката. Пусть предикат ***P*** представляет собой конъюнкцию (логическое И) ***n*** атомарных высказываний, тогда его префиксом называется предикат, состоящий из $\boldsymbol{m}\mid\boldsymbol{m}\leq\boldsymbol{n}$ первых атомарных высказываний ***P***, который дальше будет обозначаться ***prefix(P,m)***. В частном случае, когда ***m=n***, префикс предиката совпадает с самим предикатом – ***prefix(P,n)=P***.

Отношение ***prefix*** обладает следующими свойствами:

– рефлексивности $\forall\boldsymbol{P}\mid\boldsymbol{P}\equiv\boldsymbol{prefix}\left(\boldsymbol{P},\boldsymbol{n}\right)$;

– асимметричности $\forall\boldsymbol{P}_1,\boldsymbol{P}_2\mid\boldsymbol{P}_1\equiv\boldsymbol{prefix}\left(\boldsymbol{P}_2\right)\Rightarrow\neg\boldsymbol{P}_2\equiv\boldsymbol{prefix}\left(\boldsymbol{P}_1\right)$;

– транзитивности $\forall\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3\mid\boldsymbol{P}_1\equiv\boldsymbol{prefix}\left(\boldsymbol{P}_2\right)\wedge\boldsymbol{P}_2\equiv\boldsymbol{prefix}\left(\boldsymbol{P}_3\right)\Rightarrow\boldsymbol{P}_1\equiv\boldsymbol{prefix}\left(\boldsymbol{P}_3\right)$.

Определение адресного реквизита. Адресным реквизитом $\boldsymbol{R}$ множества адресов $\boldsymbol{A}$ называется свойство, присущее множеству адресов $a \in A$ так, что, либо его общий предикат $\boldsymbol{P}(\boldsymbol{R})$ полностью совпадает с контекстом $\boldsymbol{C}(\boldsymbol{A})$, либо общий предикат этого реквизита $\boldsymbol{R}$ представляет собой префикс контекста множества $\boldsymbol{A}$, т.е. $\boldsymbol{P}(\boldsymbol{R}) \equiv \boldsymbol{prefix}(\boldsymbol{C}(\boldsymbol{A}), \boldsymbol{m}) \mid \boldsymbol{m} \le \boldsymbol{n}$. Множество значений адресного реквизита состоит из уникальных, т.е. неповторяющихся значений.

Определение основного адресного реквизита. Основным адресным реквизитом $\boldsymbol{R}$ множества адресов $\boldsymbol{A}$ считается адресный реквизит, определённый на всём множестве, т.е. это свойство, характеризующее каждый адрес $a \in A$. Основной адресный реквизит – синоним ключевого реквизита, а также кандидата в ключевые реквизиты, в случае, когда множество представлено в форме адресного списка. Адресный реквизит, не являющийся основным, в дальнейшем будет называться *вспомогательным.*

Пусть заданы два значения различных адресных реквизитов $r_1 \in R_1$ и $r_2 \in R_2$, такие, что предикат значения $\boldsymbol{r}_1$ (где $\boldsymbol{r}_1$ – конъюнкция контекста значения с идентифицирующим выражением, т.е. $C(r_1) \wedge (R_2 = r_1)$) является префиксом контекста r_2, $P(r_1) \equiv prefix(C(r_2), m)$. Тогда значение r_2 совместимо со значением r_1, что будет обозначаться как $r_2 \equiv comp(r_1)$.

Отношение совместимости на множествах значений R_1 и R_2 обладает свойствами *антирефлексивности*, *асимметричности*, но оно обладает *транзитивностью*, которое наследует от транзитивности отношения ***prefix***.

Значения $r_2 \in R_2$, совместимые с $r_1 \in R_1$, образуют подмножество $R_2^{r_1} \subseteq R_2$, в результате чего R_2 разбивается на две непересекающиеся части – совместимую и несовместимую с $\boldsymbol{r}_1$.

Пусть множество R_1 такое, что для любого значения $r_1 \in R_1$ существует, по крайней мере, одно значение $r_2 \in R_2$, совместимое с $\boldsymbol{r}_1$. Предикаты различных значений из R_1 различны по определению. Следовательно, для каждой пары различных значений из R_1 соответствующие им подмножества совместимых значений не пересекаются, а множество значений $r_2 \in R_2$, совместимых хотя бы с одним значением их R_1, в общем случае является подмножеством R_2 и будет обозначаться $comp(R_2, R_1) \subseteq R_2$. Тогда, если $comp(R_2, R_1) \subset R_2$, т.е. объединение подмножеств, совместимых R_1, не совпадает с множеством значений реквизита R_2, то реквизит R_2 *полусовместим* с реквизитом R_1. Если $comp(R_2, R_1) = R_2$, то реквизит R_2 *совместим* с реквизитом R_1. Если $comp(R_2, R_1) \neq \varnothing$, т.е. множество значений R_2, совместимых со всеми значениями R_1, не пусто, то R_1 определяет отношение эквивалентности на всём множестве реквизита R_2 или на его подмножестве.

Следствие. Для каждого адресного реквизита $\boldsymbol{R}$ существует совместимое с ним подмножество адресов $\boldsymbol{A'} \subseteq \boldsymbol{A}$. При этом полное множество адресов A совместимо с каждым основным реквизитом $\boldsymbol{R}$.

То есть существенным свойством адресного реквизита, без которого он не может существовать, является наличие совместимых с ним адресов. На этой основе можно сформулировать признак совместимости адресных реквизитов.

2.3. Признак совместимости адресных реквизитов

Утверждение 1. Адресные реквизиты R_1, и R_2, совместимы тогда и только тогда, когда существует непустое множество адресов A, совместимое с каждым из них.

Доказательство.

Пусть $R_2 \equiv comp(R_1)$. Тогда $\forall r_2 \in R_2 \exists r_1 \in R_1 \mid P(r_1) \equiv prefix(C(r_2), m)$.

Кроме того, реквизит R_2 совместим с множеством адресов, т.е. $A \equiv comp(R_2)$. Тогда $\forall a \in A \ \exists r_2 \in R_2 \mid P(r_2) \equiv prefix(C(a), m)$. В силу транзитивности отношения $\forall a \in A \ \exists r_1 \in R_1 \mid P(r_1) \equiv prefix(C(a), m)$. То есть множество адресов А сравнимо с реквизитом R_1, $A \equiv comp(R_1)$.

Примечание. Далее, когда это не приводит к двусмысленности, вместо $\forall r_i \in R_1 \mid P(r_i)$ и $\forall a \in A \mid C(a)$ будут использоваться сокращённые формы записи $P(R_i), C(R_i), P(A), C(A)$).

Утверждение 2 (обратное). Если оба реквизита сравнимы с одним и тем же множеством адресов $A \equiv comp(R_1) \wedge A \equiv comp(R_2)$, тогда $P(R_1) \equiv prefix(C(A), m)$ $P(R_2) \equiv prefix(C(A), k)$, то есть предикаты значений этих реквизитов являются префиксами контекста соответствующих адресов общего множества.

В этом случае, если $m < k$, то $R_2 \equiv comp(R_1)$; если m, то $R_1 \equiv comp(R_2)$; если $m = k$, то R_1 и R_2 – один и тот же реквизит.

2.4. Отношение эквивалентности на значениях адресного реквизита

В общем случае каждое множество значений адресного реквизита $\boldsymbol{R}$ представляет собой объединение двух непересекающихся множеств: множество *нормальных* или *эталонных значений* $\boldsymbol{R^S}$ и множество *расширений эталонов* $\overline{\boldsymbol{R^S}}$. При этом каждое значение $\overline{\boldsymbol{R^S}}$ соответствует только одному значению множества эталонов, т.е. $\forall r \in \overline{\boldsymbol{R^S}} \exists ! \boldsymbol{r^S} \in \boldsymbol{R^S}$. Каждое значение множества расширений эталонов называется *синонимом* для своего эталона.

Обозначим через $\boldsymbol{E}$ функцию, которая каждому значению адресного реквизита $\boldsymbol{R}$ ставит в соответствие его эталон $E(r) = r^S$, где $r \in R \wedge r^S \in R$.

По определению все значения адресного реквизита уникальны, следовательно, уникальны как значения эталонов $\boldsymbol{r^S} \in \boldsymbol{R^S}$, так и расширенные значения $r \in \overline{R^S}$. Теперь можно уточнить высказывание контекста для множества значений основного реквизита $\boldsymbol{R_j}$:

$$\wedge_1^{i-1} \left(R_k = E_k \left(r_k \right) \right) \Rightarrow R_j .$$

То есть реквизиты контекста принимают эталонные значения. Порядок на множестве реквизитов адреса введен [1, с. 59] независимо от формы представления множества адресов.

2.5. Признак эквивалентности адресных списков

Два адресных списка A_1 и A_2 эквивалентны, если существует взаимно однозначное соответствие между предикатами элементов этих списков. Под соответствием между предикатами элементов $a_1 \in A_1$ и $a_2 \in A_2$ рассматривается выполнение следующих условий:

1) совпадение количества и названий реквизитов $\boldsymbol{R_i^N}(\boldsymbol{A_1}) = \boldsymbol{R_i^N}(\boldsymbol{A_2})$;

2) равенство множеств значений реквизитов $\boldsymbol{R_i^D}(\boldsymbol{A_1}) = \boldsymbol{R_i^D}(\boldsymbol{A_2})$;

3) равенство значений в правой части каждого атомарного выражения с точностью до эталона, т.е.

$$\forall \boldsymbol{i}: \boldsymbol{R_i^N}(\boldsymbol{a_1}) = \boldsymbol{r_i}(\boldsymbol{a_1}) \wedge \boldsymbol{R_i^N}(\boldsymbol{a_2}) = \boldsymbol{r_i}(\boldsymbol{a_2}) \rightarrow \boldsymbol{E_i}(\boldsymbol{r_i}(\boldsymbol{a_1})) = \boldsymbol{E_i}(\boldsymbol{r_i}(\boldsymbol{a_2})).$$

Следует обратить внимание, что индекс *i* идентифицирует не порядок следования значений реквизита в адресном списке, а лишь совпадение реквизитов в различных списках. Более того, название реквизита R_i^N – своего рода указатель на ось координат, а множество значений R_i^D – шкала значений на этой оси координат в векторном пространстве реквизитов, содержащем адреса. При этом третье условие признака гарантирует возможность построения ключевых списков, совпадающих как по составу реквизитов, так и по набору значений.

3. Порождающая грамматика

3.1. Постановка задачи построения порождающей грамматики

Задача этого раздела формулируется следующим образом. Пусть дан адрес в форме последовательности значений реквизитов, т.е. $a \in A$. Требуется на его основе построить адрес в строковом фор-

мате $a^s \in A^s$. В более общем виде эта задача формулируется так – построить алгоритм, переводящий любой a в соответствующий ему a^s.

Порождающая грамматика в нашем случае есть четверка

$$AG = \left(V_N, V_T, P(a), AS\right),$$

где $V_N = \left\{ \boldsymbol{R_i^N} \right\}$ – множество названий реквизитов, $V_N = R_1^D \cup \ldots R_n^D$; $V_T = \left\{ \boldsymbol{R_i^T} \right\}$ – объединённое множество значений реквизитов (далее, объединённый словарь реквизитов); i – номер реквизита; $AS \in V_N$ – начальный символ; $P(a)$ – конечное множество правил вывода или правил подстановки, которое зависит от пореквизитного вектора значений a.

Правила порождающей грамматики:

$$\begin{cases} AS \to R_1 \\ R_1 \to r_1 R_2 \\ r_1 \ldots r_{i-1} R_i \to r_1 \ldots r_{i-1} r_i R_{i+1} \mid 2 \le i < n-1 \\ r_1 \ldots r_{n-1} R_n \to r_1 \ldots r_{i-1} r_n \end{cases} \quad 1.1$$

Здесь r_i значение реквизита R_i , полученное из вектора $a = \{r_1, r_2, \ldots r_n\}$. Последнее правило называется *заключительным правилом*. При этом каждый реквизит R_{i+1} совместим с R_i , благодаря чему все незаключительные правила являются линейными, т.к. $R_i \to r_i R_{i+1}$ [8, с. 172].

Пусть φ и ψ – цепочки в составе вышеуказанных правил, таких, например, как $r_1 R_2$, $r_1 \ldots r_{i-1} R_i$, или $r_1 \ldots r_{i-1} r_n$. Тогда последовательность цепочек D=(φ_k, φ_{k+1},…, φ_{k+m}) называется φ-*выводом для* ψ, если φ= r_k, ψ= r_{k+m}, и $\exists \chi, \psi \mid_{k+j} \to \chi_{k+j+1} \psi$. В этом случае будем говорить, что ψ выводима из φ или $\varphi \Rightarrow \psi$. Из линейности грамматических правил следует $\forall i \left\langle j \middle| R_i \Rightarrow R_j \right.$.

На множестве реквизитов может быть введено отношение порядка «≤»:

1) $R_i \le R_j$ для всех $i=j$;

2) $R_i \le R_j$ для всех $i \left\langle j \middle| R_i \Rightarrow R_j \right.$.

Легко видеть, что свойства рефлексивности, транзитивности и антисимметричности[2] для введенного отношения выполняются. Так, рефлексивность отношения следует из определения. Транзитивность и антисимметричность отношения «≤» следует из транзитивности и антисимметричности операции $R_i \Rightarrow R_j$. Отношение порядка введено на всём множестве реквизитов, значит множество реквизитов – линейно упорядоченное множество, а грамматика – последовательная грамматика [8, с. 172].

Далее будем называть построенную грамматику AG эталонной адресной грамматикой или просто эталонной грамматикой, порядок на множестве реквизитов эталонной грамматики будем называть естественным порядком. Порождённые грамматикой AG адресные строки будем называть *строками с естественным порядком*, если значения реквизитов в них следуют в порядке, соответствующем порядку на множестве реквизитов.

Грамматика AG из-за своих упрощенных возможностей не представляет серьёзного практического интереса, но и в таком виде позволяет понять, какие упрощения и ограничения должны быть исключены по мере её развития. *Во-первых*, как показано на рисунке 1, грамматика AG допускает независимость значений различных реквизитов, позволяя тем самым порождать синтаксически правильные, но неверные по смыслу адреса. Для того чтобы решить эту проблему, следует учесть, что значения разных реквизитов зависимы друг от друга. *Во-вторых*, значения адресных реквизитов не являются атомарными [2, с. 47], а значит, одно и то же значение может быть представлено в разных формах. *В-третьих*, не все адресные строки являются строками с естественным порядком следования значений реквизитов. *В-четвёртых*, грамматика AG может порождать только полные адреса, которые содержат значения всех реквизитов от первого до последнего. В то же время существует потребность в порождении неполных адресных строк, таких, которые заканчиваются значением заранее заданного реквизита.

[2] *Курош А.Г.* Лекции по общей алгебре: учебник. – 2-е изд., стер. – Санкт-Петербург: Лань, 2007. – 560 с.

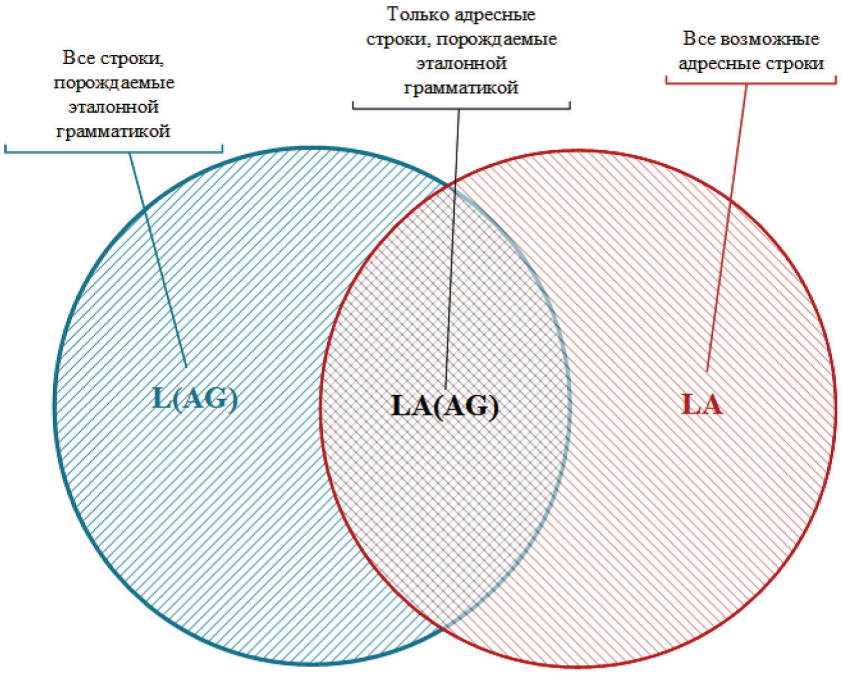

Рисунок 1 – Множества адресных строк, порождаемых грамматикой AG

3.2. Семантика словарей нетерминальных символов

Множество значений каждого реквизита адресов определяется предикатом реквизита. При этом предикаты адресных реквизитов взаимозависимы. В то же время, грамматика *GA*, будучи грамматикой непосредственных составляющих, имеет дело со значениями реквизита, независимо от их предиката. Игнорирование зависимости реквизитов от их предикатов приводит к тому, что язык, порождаемый грамматикой $L(GA)$, много шире языка актуальных адресов. В связи с этим возникает вопрос: можно ли изменить грамматику AG так, чтобы порождаемый ей язык $L(GA)$ полностью совпадал с языком актуальных адресов $LA(GA)$.

Поставленную задачу начнём с грамматического анализа языка (множества) адресов, предварительно ограничив для краткости множество значений каждого реквизита только эталонными значениями. Для этого построим дерево R_iT по следующим правилам:

– вершине в корне дерева дадим условное значение R_iS;

– проведём рёбра к вершинам, находящимся на расстоянии 1 от корня дерева, и присвоим им значения реквизита R_1;

– из каждой вершины, которая находится на расстоянии $k \mid 1 < k \le i$ от корня дерева, проведём рёбра-потомки со значениями реквизита R_k, для которых значения в родительской цепи входят в состав предиката этого реквизита.

Построенное дерево R_iT состоит из двух частей: поддерева $R_{i-1}T$ и вершин со значениями R_i, соединённых рёбрами с этим поддеревом. В то же время, реквизит по отношению к своему значению является смысловой категорией, а введенная выше грамматика – категориальной грамматикой [8, с. 221–222]. Учитывая этот факт, далее поддерево $R_{i-1}T$ будем называть *поддеревом категорий* для R_iT, т.к. категорией каждого из значений реквизита R_i является значение предыдущего реквизита, а точнее, набор всех значений в цепи от корня до *i*-й вершины. При этом, каждая полная цепь от корня до листа дерева отражает смысловое значение реквизита, т.к. последняя вершина содержит непосредственное значение реквизита, а все предыдущие вершины пути представляют собой набор расширяющихся категорий этого значения.

Отмеченное свойство позволяет рассматривать построенное дерево как некоторое префиксное дерево[3], в котором значения реквизитов мысленно сжаты до символов специального алфавита адресов.

Особенный интерес представляет дерево максимального реквизита R_nT, т.к. конкатенация (сцепление) всех значений в вершинах полного пути формирует адресную строку a^s при условии, что конкатенация выполняется с учётом заранее определённых разделителей. Следовательно, дерево максимального реквизита совпадет с деревом адресов AT.

В отличие от адресной грамматики, которая характеризует синтаксическую структуру адресных строк, полные цепи в дереве адресов показывают смысловую (семантическую) структуру адресов. Действительно, переменная $\boldsymbol{r_i}$ в дереве адресов пробегает не все значения $\boldsymbol{R_i}$, а лишь те, для которых истинным является высказывание $\wedge_1^{i-1}\left(R_k = r_k\right)$. Поэтому, только адресные строки, сформированные на основании значений полных цепей дерева адресов AT, составляют язык актуальных адресов $LA\left(GA\right)$.

Замечание

Дерево адресов AT имеет очень обобщённую, вследствие чего, упрощённую структуру. Для того чтобы составить представление о структуре реального дерева адресов, можно обратиться к статье с анализом семантической структуры адресов ФИАС [9].

Конец замечания

На рисунке 2 показана предварительная технология порождения строкового адреса.

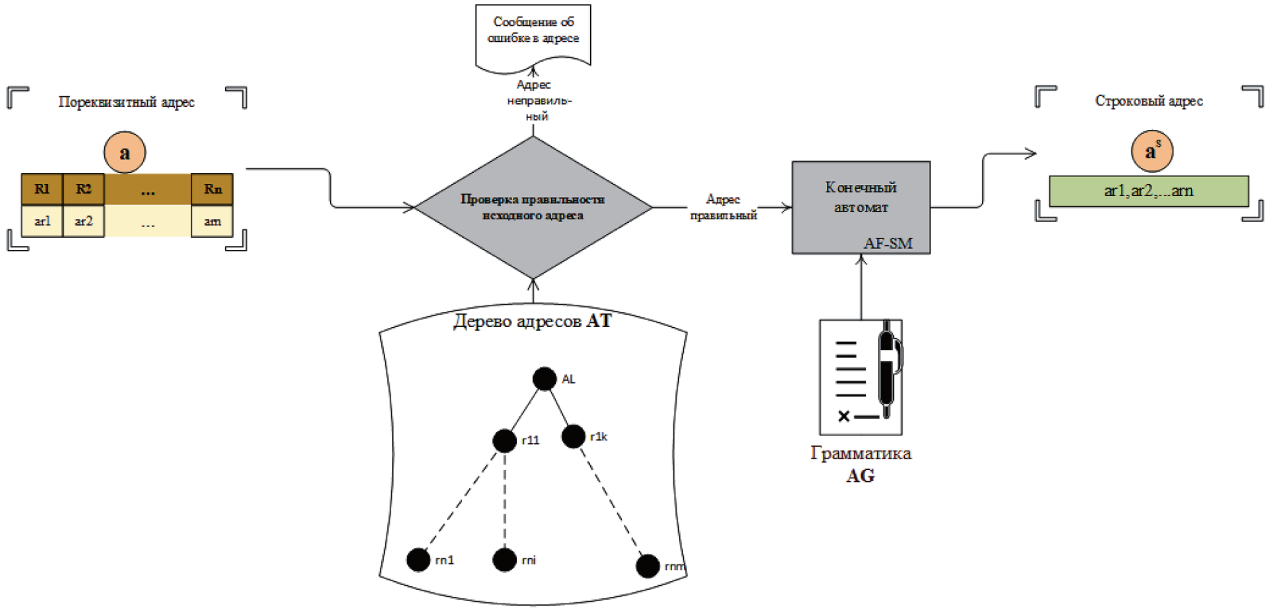

Рисунок 2 – Предварительная технология порождения строкового адреса

Теперь вопрос, заданный в начале этого раздела, может быть переформулирован так – имеет ли смысл строить грамматику, порождающую только актуальные адреса. В ответ на него анализ структуры словарей реквизитов и дерева адресов позволяет сделать следующие выводы. *Во-первых*, любая грамматика, порождающая адресные строки, должна быть контекстно-зависимой [10, с. 31], т.к. каждое следующее значение извлекается из множества, которое определено набором предыдущих значений. *Во-вторых*, конечно, можно описать выделенное множество адресов в форме набора грамматических правил. Результат такого построения следует скорее назвать ещё одним описанием исходного множе-

[3] Другие названия: бор, луч, нагруженное дерево, англ. trie.

ства, только в другой форме. Кроме того, количество правил такого описания будет огромно. Так, для того чтобы описать адреса Красноярского края, понадобится свыше 600 тысяч грамматических правил. При этом способов упрощения за счёт обобщения правил этого описания пока не видно. *В-третьих*, для адресов характерна высокая степень изменчивости, как среди значений реквизитов, так и естественного порядка их следования, что приведёт к необходимости внесения изменений не только дерева адресов, но и эталонной адресной грамматики.

Но вернёмся к построенной ранее грамматике AG. Она была определена как некий шаблон, который «превращается» в грамматику после подстановки в правилах $P(a)$ действительных значений пореквизитного адреса a на место их знаков. Если представить, что на вход грамматике AG передан пореквизитный адрес, извлечённый из дерева адресов AT, то порождённый строковый адрес окажется актуальным. То же самое можно сказать об адресе, прошедшем проверку на присутствие в дереве адресов.

Действительно, как уже отмечалось выше, дерево адресов имеет форму префиксного дерева, которое используется для поиска (распознавания) строк (ключевых слов), символы которых распределены по рёбрам дерева. С этой целью, например, применяется алгоритм Ахо-Корасик (Aho-Corasick), описание которого можно найти, например, в [11, с. 187]. Поэтому, построенное дерево адресов, как показано на рисунке 2, может быть использовано как инструмент предварительной оценки и/или преобразования пореквизитного адреса перед передачей его конечному автомату, порождающему строковый адрес по правилам грамматики AG.

3.3. Перестановки реквизитов в адресных строках

Прежде чем перейти к обсуждению реализации порождения адресных строк, в которых значения реквизитов не следуют естественному порядку, остановимся на методе подстановки в правила $P(a)$ пореквизитного адреса a с естественным порядком, анализ которого приведёт к лучшему пониманию решения основной задачи раздела.

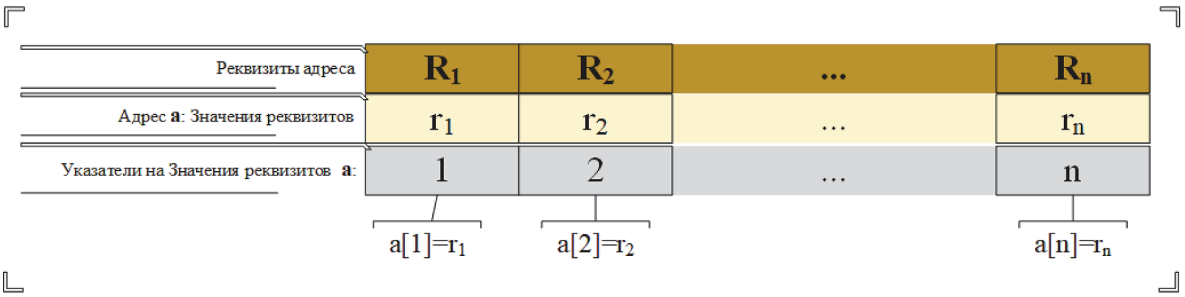

Рисунок 3 – Механизм подстановки в грамматику AG адреса с естественным порядком

Для того чтобы грамматика AG позволяла порождать адресную строку из любого заранее заданного пореквизитного адреса, необходимо в списке её правил вместо абсолютных значений реквизитов подставить указатели на место их нахождения в памяти, куда эти значения были предварительно загружены. Так, на рисунке 3 показан список *a*, содержащий значения некоторого адреса с естественным порядком следования реквизитов. Доступ к значению каждого реквизита осуществляется по его порядковому номеру, т.к. $a[i]=r_i$. Поэтому, для того чтобы обеспечить подстановку значений реквизитов адреса в правила грамматики AG, достаточно в описании 1.1 заменить каждое в вхождение r_i на $a[i]$. В результате получится следующее описание $P(a)$:

$$\begin{cases} AS \to R_1 \\ R_1 \to a[1]R_2 \\ a[1]\ldots a[i-1]R_i \to [1]\ldots a[i-1]a[i]R_{i+1} \mid 2 \le i < n-1. \\ a[1]\ldots a[n-1]R_n \to a[1]\ldots a[n-1]a[n] \end{cases} \quad 1.2$$

Так, заменяя абсолютные значения реквизитов адреса на относительные, обеспечивается дополнительная гибкость в использовании порождающих правил грамматики.

Теперь покажем, как можно решить задачу порождения адресных строк с произвольным порядком следования значений реквизитов.

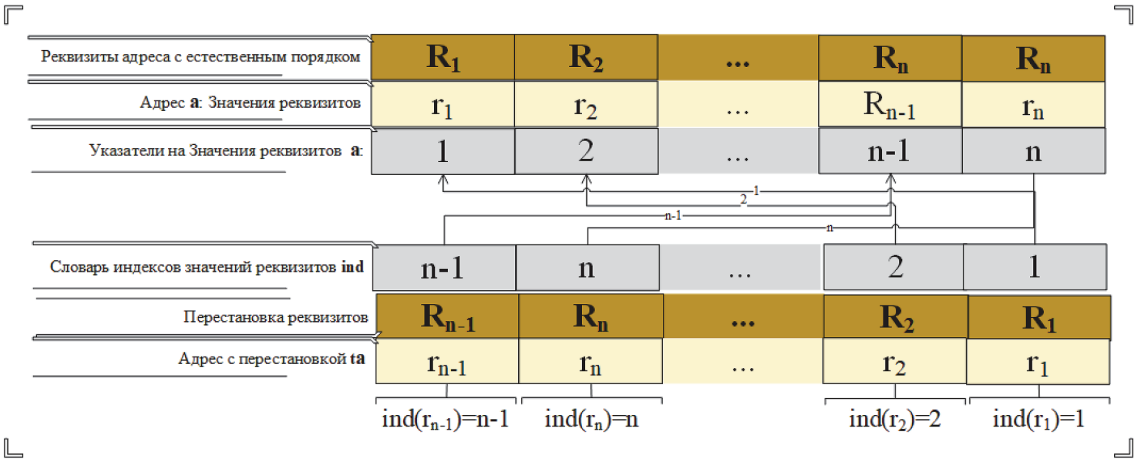

Рисунок 4 – Механизм подстановки в грамматику AG адреса с произвольным порядком

Для начала введем функцию ind , которая каждому значению реквизита ставит в соответствие его номер в естественном порядке, т.е. $ind\left(r_i\right)=i$, и назовём ta список с произвольным порядком хранения значений реквизитов. Тогда $ind\left(ta\left[i\right]\right)$ – порядковый номер значения реквизита $ta\left[i\right]$ в естественном порядке, а $a\left[ind\left[ta\left[i\right]\right]\right]$ – подстановка, которая должна быть выполнена в грамматических правилах вместо r_i .

3.4. Исключение реквизитов из адресных строк

Есть два конкурирующих принципа порождения адресных строк: максимальная информативность или минимальная длина. При этом результат применения и того, и другого не должен нарушать условий однозначности и уникальности адреса [1, с. 59–60]. В приложении к адресному дереву эти признаки требуют, чтобы для любого *i* в дереве *i-го* реквизита каждые два полных пути были различными. Отличаться они должны либо длиной пути, либо значениями в равноудалённых от корня рёбрах.

Добиться уменьшения длины полного пути дерева адресов, а, следовательно, и порождаемых адресных строк можно, если ограничиться не максимальной, а достаточной их информативностью. В этих условиях самым очевидным методом уменьшения длины порождаемых адресных строк является ликвидация *дублирования корня* в адресном дереве. Рассмотрим эту процедуру.

Дерево адресов со смысловой точки зрения определяется своим логическим предикатом, который здесь называется контекстом. При этом информация, которая явно не представлена в рёбрах дерева адресов, может быть дополнена данными контекста как некая информация умолчания. Например, адреса на территории Российской Федерации в дереве адресов могут иметь значение «Россия» в первом от корня ребре. Но это ребро будет единственным в дереве и общим для всех адресов. Поэтому, если корень такого дерева адресов перенести в вершину, находящуюся на расстоянии 1 от корня, то длина каждого полного пути уменьшится на единицу, а значение первого ребра станет частью контекста.

От этого частного случая перейдём к более общему методу *сжатия дерева адресов*, под которым понимается следующая операция.

Сжатие дерева адресов. *I-сжатием*, *полным сжатием* или просто *сжатием дерева адресов* называется операция по удалению *i*-го ребра в каждом полном пути путём соединения выходящей и исходящих вершин этого ребра.

В первую очередь интерес представляют сжатия дерева адресов, после применения которых не нарушаются ни однозначность, ни уникальность полных путей результирующего дерева адресов. По месту применения в дереве адресов из таких операций можно образовать три группы. Одна группа таких операций от корня дерева адресов описана выше на примере ликвидации дублирования корня. Вторая группа – это операции сжатия со стороны листьев. Несмотря на то, что в результате применения операций этой группы вновь образованное дерево не теряет однозначности, но его полные пути из адресов становятся реквизитами адресов в дерево справочника реквизита. Третья группа – операции сжатия к середине дерева. Такие операции удаляют ребро на расстоянии *i* от корня дерева. При этом, по построению дерева, удаляемые рёбра содержат значения *i*-го реквизита.

Описанные операции сжатия не сказываются ни на значениях в каждом из оставшихся рёбер дерева, ни на их смысловой категории. В то же время существуют формы сжатия адресных строк, реализация которых может быть осуществлена только совместно с некоторыми операциями соединения адресных реквизитов.

3.5. Операции над адресными реквизитами

Отдельные реквизиты, с одной стороны, не являются самостоятельными понятиями, т.к. представляют собой свойства адреса, с другой, адрес – это свойство объекта, позволяющее его найти. При этом поиск каждой категории объектов может осуществляться в своём особом виде пространства со своими адресными реквизитами, выполняющими роль координатных осей. Значит совокупность реквизитов является существенным свойством, определяющим понятие «адрес», но для разных категорий адресов эта совокупность может быть разной. Таким образом, адреса в каждом пространстве могут состоять из своего набора реквизитов. А раз так, то имеет смысл изучать не только реквизиты как статические свойства конкретной категории адресов, но и операции, образующие адреса из реквизитов, выделенных в отдельное множество.

Соединение реквизитов

Итак, определим множество адресных реквизитов UR, состоящее из двух групп: основные и вспомогательные. При этом реквизиты этого множества попарно могут быть совместимыми, полусовместимыми и несовместимыми.

До сих пор, как при построении дерева адресов, так и адресной грамматики неявно использовалось *строгое соединение* совместимых друг с другом основных адресных реквизитов. В результате все значения на определённом уровне дерева адресов соответствовали одному реквизиту. Значение реквизита в адресной строке при таком соединении занимает заранее определённую позицию. Конечно, если адресная строка создаётся в соответствии с эталонным порядком. Грамматика и дерево адресов с такими свойствами не слишком конструктивны.

Пусть реквизит R_2 полусовместим с реквизитом R_1. А это значит, что только часть значений R_1 образуют покрытие значений R_2 непересекающимися подмножествами. При строгом соединении создаётся состояние неопределённости для значений из R_1, для которых не нашлось совместимых значений из R_2. Для преодоления указанной неопределенности дополним правило строгого соединения, позволив на выходе дублировать значения R_1 без совместимых значений из R_2. Более точно расширенное соединение можно определить следующим образом:

$$R_1 \circ R_2 \equiv \begin{cases} r_2 \in R_2, \textit{если} \exists r_1 \in R_1 \textit{такой, что} r_2 \textit{совместим с} r_1 \\ r_1 \in R_1, \textit{если} \nexists r_2 \in R_2 \textit{такой, что} r_2 \textit{совместим с} r_1 \end{cases}.$$

Данное определение соединения охватывает оба случая, как полной совместимости, так и полусовместимости R_2 по отношению R_1. Но в силу того, что образованное множество значений уже не принадлежит одному реквизиту, для каждого его значения придется явно или неявно хранить сведения о принадлежности к родительскому или дочернему реквизиту. При этом в порождённой адресной строке двойное и более повторение значения родительского реквизита будет заменяться одним-единственным.

Исключение среднего реквизита

Рассмотрим операцию последовательного соединения трёх совместимых реквизитов – $R_1 \circ R_2 \circ R_3$. Под исключением среднего будем понимать преобразование этого соединения без потери однозначности и уникальности к виду $R_1 \circ R_3$.

Необходимым условием применения операции соединения является совместимость участвующих в ней реквизитов. А так как отношение совместимости транзитивно, то возможность операции исключения среднего оправдана. Сложность лишь в том, что удаление реквизита из соединения может изменить однозначность его значений.

Действительно, операция соединения добавляет одноместные предикаты, содержащие значения реквизитов R_1 и R_2, в контекст значений R_3. Операция же исключения R_2 удаляет одноместный предикат, содержащий значения R_2, из контекста значений R_3. В результате такого изменения значения соединения $R_1 \circ R_3$ могут оказаться неоднозначными. Причина в том, что два одинаковых значения реквизита R_3 отличались контекстом, но после удаления из контекста одноместного предиката эти значения могут стать неразличимыми, что приведёт к нарушению принципа однозначности адреса. Обеспечить однозначность результата исключения среднего можно путем введения дополнительных операций над соединяемыми реквизитами, главной из которых станет операция их свёртки, которая из двух множеств значений реквизитов R_2 и R_3 создаёт одно множество соединённых значений.

Определим свёртку $R_2 \circ R_3$ следующим образом:

$$R_3\left(R_2\right) \equiv \begin{cases} r_3\left(r_2\right) \mid r_3 \in R_3 \wedge \exists r_2 \in R_2 \ \wedge \ r_3 \equiv comp\left(r_2\right). \\ r_2 \mid \nexists r_3 \in R_3 \ \wedge \ r_3 \equiv comp\left(r_2\right) \end{cases}$$

Формула $r_3 \equiv comp\left(r_2\right)$ в определении свертки указывает на то, что r_2 совместим с r_3, в целом свёртка приводит к дополнению значения r_3 значением r_2, тем самым предотвращая возникновение возможной неопределённости. Соединённые значения при этом выглядят, например, так: «посёлок (район)», «улица (село)», и т.д.

С одной стороны, свёртка $R_3\left(R_2\right)$ обеспечивает однозначность для потенциально повторяющихся значений реквизита R_3, с другой – создает избыточность в виде соединённых значений для неповторяющихся значений этого реквизита.

Действительно, основное отличие соединённого от однородного значения состоит в том, что его предикат получается в результате конъюнкции его контекста не с одним, а с двумя одноместными предикатами, т.е. $P\left(r_3\left(r_2\right)\right) \equiv C\left(r_2\right) \wedge \left(R_2 = r_2\right) \wedge \left(R_3 = r_3\right)$. При этом, значение r_3 неповторимо тогда и только тогда, когда множество адресов с префиксом $P\left(r_3\left(r_2\right)\right)$ остается неизменным при удалении из его префикса предиката $\left(R_2 = r_2\right)$.

Для преодоления избыточности соединенных значений следует применить вспомогательную операцию сжатия соединённого значения, которая преобразует $r_3\left(r_2\right)$ к r_3, если это значение неповторимо, и оставляет $r_3\left(r_2\right)$ неизменным в противном случае. Операцию сжатия можно записать через отношение совместимости.

Сжатие $R_3\left(R_2\right)$ можно представить в виде

$$reduce(r_3\left(r_2\right)) \equiv \begin{cases} r_3\left(r_2\right) \mid \left\{comp\left(a, r_3\left(r_2\right)\right)\right\} \supset \left\{comp\left(a, r_3\right)\right\} \\ r_3 \mid \left\{comp\left(a, r_3\left(r_2\right)\right)\right\} = \left\{comp\left(a, r_3\right)\right\} \end{cases}.$$

Другими словами, если множество совместимых с r_3 адресом осталось неизменным после свёртки с r_2, то сжатие исключает из соединённого значения дополнение $\left(r_2\right)$, иначе оставляет его без изменения.

Операций свёртки и сжатия достаточно для того, чтобы дать следующее определение исключению среднего:

$$EA\left(R_1 \circ R_2 \circ R_3\right) \equiv ExceptAverage\left(R_1 \circ R_2 \circ R_3\right) \equiv R_1 \circ reduce\left(R_3\left(R_2\right)\right).$$

Если R_1, R_2 и R_3 – последовательные рёбра дерева адресов, то исключение среднего – ещё один шаг сжатия дерева адресов, которое можно назвать сжатием с дополнением.

Отклонения, приводимые к норме

Рассмотренные выше правила образования дерева адресов и соединения адресных реквизитов образуют нормальные условия для создания строковых адресов. На практике же встречаются отклонения от этих условий, часть из которых могут быть приведены к норме рассматриваемыми здесь методами.

Наиболее вероятное отклонение от нормальных условий создания строковых адресов – наличие двух и более одинаковых значений адресного реквизита с общим контекстом. Приведение этого отклонения к норме выполнятся при помощи подбора дополнительного реквизита.

Пусть значения реквизита $r_1, r_2 \in R$ – неразличимы, т.е. имеют совпадающие значения и общий контекст. Для того чтобы преодолеть неоднозначность $r_1, r_2 \in R$, выполняется поиск адресного реквизита R_0, с которым, во-первых, совместим R, во-вторых, значения r_1, r_2 совместимы с различными значениями R_0, т.е.

$$r_{0,1}, r_{0,2} \in R_0 \mid r_{0,1} \neq r_{0,2} \wedge r_1 \equiv comp\left(r_{0,1}\right) \wedge r_2 \equiv comp\left(r_{0,2}\right).$$

Отмеченное свойство адресного реквизита R_0 делает его *различительным* для одинаковых значений реквизита R.

Соединение $R_0 \circ R$ приводит к изменению контекстов значений r_1, r_2 путём присоединения к ним атомарных предикатов $R_0 = r_{0,1}$ и $R_0 = r_{0,2}$ соответственно. А значит, результат соединения исправил изначальное отклонение.

Если изначальное число неразличимых значений невелико в сравнении со всем множеством значений R, то вместо соединения реквизитов следует использовать свертку $R\left(R_0\right)$.

Ещё одна группа отклонений возникает при соединении полусовместимых адресных реквизитов.

Действительно, если реквизит R_2 совместим с реквизитом R_1, то, как следует из определения, любое значение $r_2 \in R_2$ совместимо только с одним значением $r_1 \in R_1$. При полусовместимости это условие не соблюдается, т.к. одно значение $r_2 \in R_2$ совместимо с множеством значений $\left\{r_{1j}\right\} \subset R_1$. В этом случае r_2 может быть представлено множеством значений свёртки $\left\{r_2\left(r_{1j}\right)\right\} \subset R_2\left(R_1\right)$. Дополнительное разделение значений подчинённого реквизита на части приводит результат соединения полусовместимых реквизитов к норме, поэтому описанная операция будет называться *разделяющее соединение* и обозначаться $R_1 \bigcap R_2$.

Символ операции пересечения множеств для обозначения операции разделяющего соединения использован не случайно. Каждое значение любого адресного реквизита представляет собой подмножество универсального множества адресов. Следовательно, операция разделения значения реквизита на части – это именование новых подмножеств адресов, возникших в результате разделения множества адресов, соответствующего реквизиту r_2.

Наличие адресов с отклонениями от идеальной структуры адресного дерева – признак нарушения качества адресной системы, основанной на множестве адресов.

Введенные операции над реквизитами, с одной стороны, позволяют привести отклонения в адресах к норме, не противоречащей синтаксическим требованиям дерева адресов. С другой стороны, в результате этих операций создаются значения, представляющие собой «конструкции с гнездованием или самовосстановлением» [12, с. 15], а значит, не могут быть порождены грамматикой с непосредственными составляющими. Такие адреса из-за большей громоздкости, трудности понимания представляют собой признак снижения качества адресной системы.

Форматирование строк адресных реквизитов

В большинстве случаев значение реквизита представляется парой: типом значения реквизита и непосредственным значением внутри подмножества, соответствующего этому типу $r = \{tr, vr\}, r \in R,$ где непосредственное значение [1]. Если бы к строковой форме необходимо было бы преобразовывать только такие реквизиты, то решить такую задачу можно было бы добавлением нескольких правил к описанной выше грамматике AG. Но в общем случае задача преобразования в строку значений рекви-

зитов не столь проста. Во-первых, строковые значения реквизита имеют различную форму представления, например, полную и сокращенную. Во-вторых, значением реквизита может оказаться не одна пара, а группа пар (достаточно вспомнить номера дома, корпуса и строения). В-третьих, в исключительных случаях, к обычной паре, составляющей значение реквизита, может добавиться третий – код значения реквизита. В-четвёртых, значением может оказаться результат соединения нескольких реквизитов.

Замечание

Сложные значения, полученные в результате соединения нескольких реквизитов, чаще всего не могут быть представлены грамматикой непосредственных составляющих, поэтому они не могут быть порождены грамматикой AG, расширенной дополнительными правилами.

Эта особенность сложных значений создает трудности при решении обратной задачи – преобразования строкового адреса к его пореквизитной форме. Планируется, что такая задача будет рассмотрена в одной их следующих публикаций.

Конец замечания

Все эти рассуждения приводят к тому, что порождение строковых значений реквизитов далее будет рассматриваться как двухуровневый процесс (алгоритм). На первом уровне конечный автомат, соответствующий грамматике AG, определяет место для значения реквизита в общей адресной строке. На втором уровне выполняется функция непосредственного формирования строкового значения реквизита. Функция формирования строковых значений в своей основе использует механизм шаблонов строковых значений реквизитов. Шаблон – строка, содержащая «замещающие поля» в специальном формате. Всё множество шаблонов разделено на группы, каждая из которых является свойством своего реквизита. Общим для всех шаблонов является формат метки, а также механизм их распознавания. Замещающее поле или метка представляет собой текст в специальном формате, ограниченный скобками “{%” и “%}”:

{%[*R*.]element_name [(*n*)] [:*f*] %},

где [R.]element_name принимает значения “tr” – тип значения реквизита, “vr” – непосредственное значение реквизита, “vc” – код непосредственного значения реквизита; *R* – название реквизита, может опускаться в случае, когда шаблон относится к этому реквизиту; *n* – порядковый номер значения в группе; *f* – форма представления в строке (“full” – полная, “brief” – краткая).

Вызов функции формирования строковых значений имеет традиционный синтаксис: *r.format*(*N*|«шаблон»), где *r* – значение реквизита, *N* – номер предустановленного шаблона, «шаблон» – непосредственная строка шаблона.

Заключение

Построена формальная модель порождения из множества пореквизитных адресов подмножества адресов строковых. Построенное подмножество строковых адресов эквивалентно исходному по определению, т.к. соответствует одному и тому же адресному списку. В то же время оно является лишь частью полного множества строковых адресов. Причина этого в том, что пореквизитное множество ограничено фиксированным набором реквизитов и порождённое им подмножество строковых адресов наследует это ограничение. Отдельные адреса полного строкового множества могут содержать в себе значения еще не выявленных реквизитов.

По результатам исследования получены следующие результаты:

− предложена теоретическая модель автоматического преобразования пореквизитного нормализованного адреса в строковую форму;

− разработана методика порождающей грамматики для преобразования пореквизитного адреса в лингвистическую переменную;

− сформулирован критерий эквивалентности множеств порождаемых адресных строк;

− установлено, что структурным описанием адресной строки является путь в дереве адресов, показывающем, какой адрес для неё является идеальным для говорящего-слышащего его человека;

– установлены причины отклонения структур адресов от их идеальной формы, определены методы их преодоления.

В дальнейшем необходимо приступить к решению обратной задачи – построить структуру адресной строки, то есть выделить непосредственные составляющие и сопоставить их реквизиту. В упрощённом виде эта задача будет решаться при заданном наборе реквизитов. Расширенный вариант этой задачи будет предполагать поиск заранее несуществующих реквизитов. В качестве базовых алгоритмов предполагается использовать анализ текстов, машинное обучение, распознавание образов.

Список литературы

1. *Гладков С.Л.* Формализация понятия «адрес» // Информатизация и связь. – 2018. – № 5. – С. 57–61.
2. *Гладков С.Л.* Нормализация адреса // Информатизация и связь. – 2018. – № 5. – С. 46–50.
3. *Заде Л.А.* Понятие лингвистической переменной и его применение к принятию приближенных решений / пер. с англ. Н.И. Ринго; под ред. Н.Н. Моисеева и С.А. Орловского. – Москва: Мир, 1976. – 165 с.
4. *Гладков С.Л.* Формальные свойства совместимости списков // Образовательные ресурсы и технологии. – 2021. – № 3 (36). – С. 60–71. – DOI 10.21777/2500-2112-2021-3-60-71.
5. *Гладков С.Л.* Классификаторы и совместимость управленческих списков // Образовательные ресурсы и технологии. – 2022. – № 2 (39). – С. 49–62. – DOI 10.21777/2500-2112-2022-2-49-62.
6. *Шрейдер Ю.А.* Равенство, сходство, порядок: Популярное введение в теорию бинарных отношений. С примерами из математической лингвистики / ред. Ю.А. Шиханович. – 2-е изд. – Москва: Ленанд, 2021. – 256 с.
7. *Войшвилло Е.К.* Понятие как форма мышления: логико-гносеологический анализ. – Москва: Либроком, 2019. – 238 с.
8. *Хомский Н.* Формальные свойства грамматик // Кибернетический сборник. Новая серия. Выпуск 2. Сборник переводов / под ред. А.А. Ляпунова и О.Б. Лупанова. – Москва: Мир, 1963. – С. 121–230.
9. *Гладков С.Л.* Опыт построения модели данных предметной области информационной системы управления государственной собственностью. Часть 2 // Образовательные ресурсы и технологии. – 2023. – № 1 (42). – С. 62–81.
10. *Хомский Н., Миллер Дж.* Введение в формальный анализ естественных языков. – Москва: Либроком, 2010. – 66 с.
11. *Ахо А.В., Лам М.С., Сети Р., Ульман Дж.Д.* Компиляторы: принципы, технологии и инструментарий: пер. с англ. – 2-е изд. – Москва: Диалектика: Вильямс, 2017. – 1184 с.
12. *Хомский Н.* Аспекты теории синтаксиса / пер. с англ. А.Е. Кибрика, В.В. Раскина, Е.Ш. Шовкуна; под общ. ред. В.А. Звегинцева. – Москва: Изд-во Московского университета, 1972. – 259 с.

References

1. *Gladkov S.L.* Formalizaciya ponyatiya «adres» // Informatizaciya i svyaz’. – 2018. – № 5. – S. 57–61.
2. *Gladkov S.L.* Normalizaciya adresa // Informatizaciya i svyaz’. – 2018. – № 5. – S. 46–50.
3. *Zade L.A.* Ponyatie lingvisticheskoj peremennoj i ego primenenie k prinyatiyu priblizhennyh reshenij / per. s angl. N.I. Ringo; pod red. N.N. Moiseeva i S.A. Orlovskogo. – Moskva: Mir, 1976. – 165 s.
4. *Gladkov S.L.* Formal’nye svojstva sovmestimosti spiskov // Obrazovatel’nye resursy i tekhnologii. – 2021. – № 3 (36). – S. 60–71. – DOI 10.21777/2500-2112-2021-3-60-71.
5. *Gladkov S.L.* Klassifikatory i sovmestimost’ upravlencheskih spiskov // Obrazovatel’nye resursy i tekhnologii. – 2022. – № 2 (39). – S. 49–62. – DOI 10.21777/2500-2112-2022-2-49-62.
6. *Shrejder Yu.A.* Ravenstvo, skhodstvo, poryadok: Populyarnoe vvedenie v teoriyu binarnyh otnoshenij. S primerami iz matematicheskoj lingvistiki / red. Yu.A. Shihanovich. – 2-e izd. – Moskva: Lenand, 2021. – 256 s.
7. *Vojshvillo E.K.* Ponyatie kak forma myshleniya: logiko-gnoseologicheskij analiz. – Moskva: Librokom, 2019. – 238 s.
8. *Homskij N.* Formal’nye svojstva grammatik // Kiberneticheskij sbornik. Novaya seriya. Vypusk 2. Sbornik perevodov / pod red. A.A. Lyapunova i O.B. Lupanova. – Moskva: Mir, 1963. – S. 121–230.

9. *Gladkov S.L.* Opyt postroeniya modeli dannyh predmetnoj oblasti informacionnoj sistemy upravleniya gosudarstvennoj sobstvennost'yu. Chast' 2 // Obrazovatel'nye resursy i tekhnologii. – 2023. – № 1 (42). – S. 62–81.
10. *Homskij N., Miller Dzh*. Vvedenie v formal'nyj analiz estestvennyh yazykov. – Moskva: Librokom, 2010. – 66 s.
11. *Aho A.V., Lam M.S., Seti R., Ul'man Dzh.D.* Kompilyatory: principy, tekhnologii i instrumentarij: per. s angl. – 2-e izd. – Moskva: Dialektika: Vil'yams, 2017. – 1184 s.
12. *Homskij N.* Aspekty teorii sintaksisa / per. s angl. A.E. Kibrika, V.V. Raskina, E.Sh. Shovkuna; pod obshch. red. V.A. Zveginceva. – Moskva: Izd-vo Moskovskogo universiteta, 1972. – 259 s.