МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)
1981 год Выпуск 91

УДК 655.25+801.4

ИНДУКТИВНАЯ РЕКОНСТРУКЦИЯ ГРАММАТИК ФЛЕКТИВНЫХ ЯЗЫКОВ

М.К.Тимофеева

I. В задачах, предполагающих автоматическую обработку текстов на естественном языке, обычно возникает проблема создания формальной модели грамматики этого языка. Формализация грамматик, разработанных традиционной лингвистикой, встречает определенные трудности, связанные с их синтаксической неполнотой, обусловленной использованием семантических определений. Поэтому в прикладных системах применяются редуцированные формы этих грамматик, представляющие собой их синтаксически полные подсистемы или дополняющиеся до таковых путем замены семантических определений синтаксическими. Такая замена обычно осуществляется посредством индуктивных обобщений. Способы обобщений, используемые в разных ситуациях, обладают значительным сходством, что позволяет попытаться разработать автоматизированный метод индуктивной реконструкции грамматик, порождающий те или иные их фрагменты на основе е д и н о й методики анализа текстов и использующий единые программные средства для формализации разных фрагментов грамматик.

Разработка такого метода состоит в решении комплекса взаимосвязанных проблем: I) выбора и с х о д н о й г р а м м а т и к и – априорных предположений, служащих основой индуктивных обобщений при изучении языков рассматриваемого типа и отражающих наиболее общие синтаксические свойства этих языков; 2) создания алгоритма индуктивных обобщений, находящего р е з у л ь т и р у ю щ у ю г р а м м а т и к у – конкретную форму проявления исходной грамматики в произвольном заданном тексте (рис.I: → – информационные связи, ➙ – управление); 3) создания программной системы, конструирующей результирующую грамматику; 4) представления результирующей грамматики в виде программы синтаксического анализа.

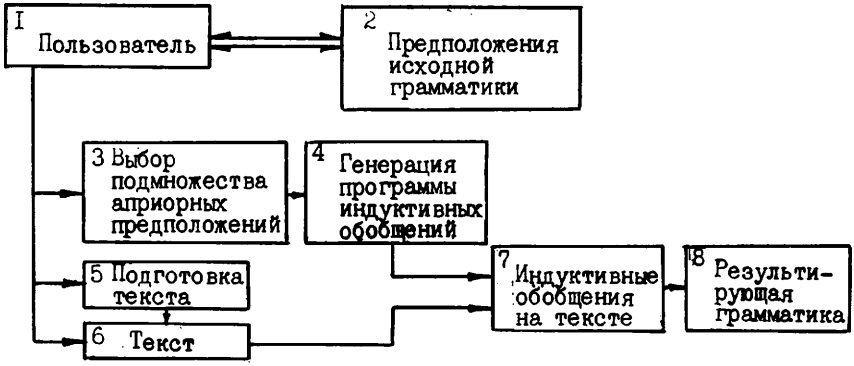

Рис. I

В [I] предложен индуктивный метод выявления парадигматических отношений текста, ориентированный на языки флективного типа (языки, например, русский, выражающие грамматические отношения через изменение слов). В данной работе предлагается расширение этого метода на область синтагматических отношений текста. По сравнению с [I] рассматриваемая в данной работе исходная грамматика обладает способностью настройки на нужное подмножество априорных предположений; содержит предположения, ускоряющие перебор подцепочек текста при выделении грамматических единиц; допускает возможность предварительной подготовки анализируемого текста.

2. Исходная грамматика должна основываться на наиболее глубоких свойствах грамматик известных языков. Одним из таких свойств является системность (см., например, [2]). При системном подходе рассматриваются как субстанциональные, так и структурные свойства грамматических единиц. Каждая единица грамматики существует в ней как структурная единица лишь постольку, поскольку она вступает в определенные отношения с другими единицами. Имеется два типа таких отношений: парадигматические и синтагматические. Парадигматические отношения объединяют единицы, противопоставленные друг другу в грамматической системе, синтагматические - единицы, связанные в линейной последовательности (тексте). Синтагматические отношения могут связывать как смежные, так и не смежные вхождения единиц в текст (слитные и разрывные синтагмы). Последовательности грамматических единиц, участвуя в системе парадигматических отношений, образуют синтагмы, т.е. парадигматические и синтагматиче-

ские отношения рассматриваются в тесной взаимосвязи: парадигматические отношения каждого следующего языкового уровня связывают синтагмы, образованные элементами предыдущего уровня.

Формы проявления указанных двух отношений на разных языковых уровнях различны. В морфологии парадигматические отношения объединяют окончания, служащие образцом формообразования некоторой части речи, или формы изменения некоторого слова. В синтаксисе рассматриваемых языков, вследствие многомерности и разнотипности характеристик синтаксических единиц, могут быть выбраны различные основы для их объединения в парадигмы (соотнесенность с одной и той же внеязыковой ситуацией, противопоставление по цели высказывания, по модально-временным характеристикам и т.д.) [3]. Т.е. чем выше уровень языковой системы, тем большее разнообразие типов парадигматических отношений на нем присутствует. То же самое верно и для синтагматических отношений.

Основной формальный признак, используемый для выделения грамматических единиц, связанных парадигматическими отношениями, - возможность их появления в одинаковых или похожих контекстах (взаимозамещаемость)*). Выделенные таким способом единицы могут интерпретироваться как синтагмы. Другим формальным признаком синтагматических отношений служит регулярная встречаемость единиц в пределах одних и тех же заданных участков текста (например, в пределах одного предложения [4]).

Основой индуктивных обобщений при изучении языка служат дистрибутивные и статистические характеристики текста. Методы, анализирующие только свойства единиц в линейной последовательности (информативность, статистическую устойчивость, отклонения от безусловной вероятности встречаемости и т.д.) и не затрагивающие область парадигматических отношений текста в большинстве случаев автоматизированы. Наиболее полный из них [4] охватывает почти все уровни языковой системы.

Автоматизация существующих методов, анализирующих как синтагматические, так и парадигматические отношения текста, встречает определенные трудности. Так, автоматизация теоретико-множественных моделей, исследующих свойство взаимозамещаемости подцепочек текста и предназначенных для описания всех языковых уровней (на -

---

*) Методы изучения языка, основанные на исследовании распределения (дистрибуции) контекстов, в которых может встречаться каждая единица, называют д и с т р и б у т и в н ы м и.

пример, [5-7]), в чистом виде практически не осуществима на имеющейся вычислительной технике, так как требует анализа всех контекстов всех подцепочек текста. Человек ускоряет этот процесс обращением к семантике текста.

Модель [8] (также не автоматизированная), исследующая как дистрибутивные, так и статистические свойства текста, существенно опирается на средства, привязанные к конкретным уровням языковой системы и различающиеся в зависимости от этого уровня. При этом средства, используемые для выявления окончаний, достаточно универсальны и могли бы быть полезны на других уровнях языковой системы, но используют определенные ограничения на позицию и длину выделяемых единиц. При снятии же этих ограничений метод становится очень трудоемким (требуется подсчет условных и безусловных вероятностей встречаемости подцепочек текста) и практически не реализуемым.

3. Основные особенности предлагаемого автоматизированного метода индуктивной реконструкции грамматик, основанного на системном подходе к языку, состоят в следующем:

1) выявление грамматических единиц основано на исследовании свойства в з а и м о з а м е щ а е м о с т и подцепочек текста;

2) парадигматическое отношение рассматривается как отношение т о л е р а н т н о с т и, а не как эквивалентность (может не выполняться свойство транзитивности [9]): требование взаимозамещаемости единиц, связанных парадигматическим отношением, во всех контекстах их встречаемости заменяется требованием их взаимозамещаемости в заданном числе контекстов (это объясняется тем, что анализируемый текст конечен и не все элементы парадигмы могут проявиться в нем одинаково ярко);

3) принятый системный подход к языку не исключает возможности анализа только линейных связей между грамматическими единицами; так, выявление синтагматических связей может осуществляться и вне анализа парадигматических отношений на основе исследования регулярности сочетаний (разрывных или слитных) между грамматическими единицами в тексте;

4) в исходную грамматику включены предположения (отсутствующие в [5-7]), использующие частотные характеристики текста и предназначенные для ускорения процесса анализа неструктурированного текста;

5) в общем случае не требуется никаких сведений об анализируемом тексте, но при наличии таковых допускаются изменения процесса его анализа путем выбора подмножества априорных предположений и специальной подготовки текста (рис. I, блоки 3 и 5).

6) принимаются во внимание не все грамматически значимые дистрибутивные свойства подцепочек текста, а, как и в [5-7], не рассматриваются контексты, не смежные с ними или вообще не являющиеся подцепочками текста; не анализируются ассоциативные связи между контекстами, более сложные, чем отношение равенства.

4. Введем некоторые определения исходной грамматики.

Пусть задан некоторый алфавит A , через A^* обозначим множество всех конечных цепочек, состоящих из символов A; Λ – пустая цепочка, $A^+ = A^* \setminus \{\Lambda\}$. Рассмотрим текст T в алфавите A, $T \in A^*$. Пусть цепочка $w = v_1 w_0 v_2 \in A^*$. Цепочку $v_1 * w_0 * v_2$, где $* \notin A$, назовем в х о ж д е н и е м цепочки w_0 в цепочку w .

Рассмотрим некоторое подмножество W подцепочек текста T. Цепочку Δ назовем л е в ы м к о н т е к с т о м вхождения $w_1 * f * w_2$ подцепочки f в T , если $w_1 = w_1'\Delta$, $w_1' \in A^*$.Подцепочки f, g текста T назовем в з а и м о з а м е щ а е м ы м и в контексте Δ , если Δ является одновременно контекстом некоторого вхождения f и некоторого вхождения g в T . Через p(w) обозначим частоту появления w в элементах W . Рассмотрим произвольную цепочку $w = a_1 \dots a_m$ и контекст $\Delta \neq \Lambda$ некоторого ее вхождения в T, такие что $\Delta w \in W$. Цепочка w в контексте Δ являет ся з в е н о м, если $m \geq 2$ и при выполнении условий 1) $p(\Delta) \neq p(\Delta a_1)$; 2) если для некоторого $i = r$ $(1 < r \leq m)$ $p(\Delta a_1 \dots a_{r-1}) = p(\Delta a_1 \dots a_r)$, то это же соотношение выполняется и для i, больших r $(i \leq m)$; 3) не существует такой цепочки $w' \neq w$, являющейся подцепочкой Δw и содержащей w , для которой выполняются первые два свойства.

Пусть задано некоторое множество P и бинарное отношение τ на этом множестве. Отношение τ называется т о л е р а н т н о с т ь ю, если оно рефлексивно и симметрично. Множество P с заданным на нем отношением толерантности τ называется п р о с т р а н с т в о м т о л е р а н т н о с т и (обозначается $\langle P, \tau \rangle$). Подмножество $P_i \subseteq P$ называется к л а с с о м т о -

л е р а н т н о с т и, если любые два элемента P_i связаны отношением толерантности τ и не существует такого элемента f ($f \in P$, $f \notin P_i$), который связан отношением τ со всеми элементами из P_i.

Предположения исходной грамматики.

П1. Любое парадигматическое или синтагматическое отношение τ проявляется в тексте не менее S раз, $S \geq 2$ (S - некоторая заданная константа, стандартное значение S равно 2). Для парадигматических отношений это означает взаимозамещаемость единиц не менее чем в S контекстах, для синтагматических - частоту встречаемости синтагм, не меньшую S.

П2. Для каждой грамматической единицы найдется не меньше S контекстов взаимозамещаемости Δ таких, что для любого из них существует цепочка w, являющаяся в этом контексте звеном.

П3. В любом классе толерантности P_i пространства $\langle E, \tau \rangle$ (E - множество грамматических единиц, τ - парадигматическое отношение) найдется хотя бы две цепочки, первые символы которых различны.

П4. Среди контекстов взаимозамещаемости любой грамматической единицы имеется хотя бы два контекста, последние символы которых различны.

П5. Разделители текста (пробел и знаки препинания) чаще всех остальных символов алфавита появляются в позиции, предшествующей линейному участку длины, большей k в деревьях 1-грамм*), соответствующих этому тексту (k - некоторая заданная константа, $k \geq 2$, по умолчанию k равно 4).

Предположение 2 принято для ускорения дистрибутивного анализа. На значимость этого свойства для грамматических единиц указывалось в [10], проверка его для окончаний слов русского языка дала положительные результаты (которые частично приведены в [1]). Предположения, подобные П3, П4, используются для выделения окончаний слов в [8]. Предположение 5 было проверено на техническом тексте русского языка (см. таблицу). Распределение частот симво -

Таблица

␣ - 207	О - 31	А - 20	Е - 19	Т - 12	Р - 7	Л - 6	У - 4	К - 4 Г-2
, - 105	П - 27	Н - 20	С - 13	И - 8	Х - 7	Ч - 6	Д - 4	Э - 3 Ж-1
. - 32	Й - 27	В - 19	Я - 12	Ы - 8	М - 7	З - 5	Ь - 4	Ю - 2

*) 1-грамма - цепочка из 1 символов; алгоритм набора статистики 1-грамм для произвольного текста приведен в [11]; способ упаковки 1-грамм в деревья описан в [1].

лов в указанной позиции получено в результате анализа одного дерева 1-грамм объемом около 11000 вершин. Предположение 5 может быть использовано при задании анализируемого множества подцепочек W текста T. Допускается два способа задания этого множества: список цепочек или перечень символов, разделяющих анализируемые цепочки в тексте (например, любое подмножество разделителей текста).

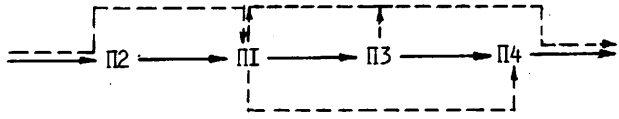

Рис. 2

Каждое из ПI–П4 реализуется в виде отдельного программного модуля, что позволяет выбирать некоторые подмножества предположений (рис.2: → - переход по умолчанию, -→ - переход, задаваемый пользователем). Анализ правых контекстов может осуществляться тем же методом при предварительном инвертировании текста.

5. Введем определение результирующей грамматики. На основе выбранного множества априорных предположений $\mathcal{A}$ генерируется программа индуктивных обобщений $\mathcal{M}(\mathcal{A})$, которая сопоставляет заданному тексту T результирующую грамматику G_T, состоящую в общем случае из трех объектов: $G_T = \langle \pi, Q, Z \rangle$, где $\pi = \langle E, \tau \rangle$ - парадигматическая структура, E - множество грамматических единиц (синтагм), τ - парадигматическое отношение на E ; Q - множество пар вида $\langle Q_y, y \rangle$, где y - код класса толерантности из π, Q_y - множество всех таких контекстов вхождений элементов E в T, по которым эти вхождения могут быть однозначно поставлены в соответствие y-му классу из π ; Z - множество регулярных сочетаний (разрывных или слитных) грамматических единиц текста T.

Пусть $P = \{P_1, \ldots, P_n\}$ - множество всех классов пространства $\pi = \langle E, \tau \rangle$. Выделим в E подмножество E_d, состоящее из всех $e \in E$, входящих в единственный класс из π. Для каждого $e \in E_d$ рассмотрим множество его контекстов взаимозамещаемости и выделим в нем подмножество K(e), состоящее из всех тех контекстов, которые не принадлежат множеству контекстов взаимозамещаемости никакого другого $e' \neq e$, $e' \in E_d$. В множество Q включаются пары вида $\langle \bigcup_{e \in P_y} K(e), y \rangle$, $y = 1, \ldots, n$. Множество Z составляется из всех сочетаний грамматических единиц более S раз встречающихся в

пределах заданных участков текста. Единицы внутри группы могут быть упорядочены.

Пусть тексту T сопоставлена результирующая грамматика $G_T = \langle \pi, Q, Z \rangle$. Рассмотрим текст $T' \in A^*$. Покрытие текста T' единицами грамматики G_T есть новый текст T'_1, получаемый в результате перекодировки таких вхождений цепочек $e \in E$ в T', которым предшествует контекст, содержащийся в одном из множеств Q_y и $e \in P_y$. Перекодировка текста, осуществляемая с помощью программ [12], может быть одного из следующих типов: ⟨контекст⟩ + ⟨грамматическая единица⟩ → ⟨код грамматической единицы⟩, ⟨контекст⟩ + ⟨грамматическая единица⟩ → ⟨код класса грамматической единицы⟩.

Покрытие максимально, если на основе той же грамматики G_T нельзя построить еще одно покрытие, отличное от T'_1. Результирующая грамматика G_T сопоставляет тексту T' такое максимальное покрытие, которой получается при анализе текста слева направо (выделяется самое левое вхождение $e \in E$, затем – ближайшее к нему и т.д.).

Обозначим через W' некоторое множество подцепочек текста T'. Множество W' согласовано с множеством сочетаний Z, если каждая пара единиц, содержащаяся в одной и той же цепочке $w \in W'$, входит хотя бы в одно сочетание $z \in Z$ (при этом может учитываться порядок элементов в сочетании). Текст T' согласован с Z, если для него задано множество W', согласованное с Z и каждая подцепочка T' содержится в некотором элементе W'.

Программа $\mathfrak{M}$ может работать итеративно. Перед каждой следующей итерацией анализируемый текст заменяется покрытием одного из указанных типов. При построении результирующей грамматики используются средства автоматической генерации программ [12], позволяющие получать G_T сразу в виде программы синтаксического анализа.

6. Основные типы задач, решаемых с помощью предлагаемого метода: 1) выявление подцепочек текста, связанных парадигматическими отношениями; 2) выявление синтагм (разрывных или слитных); 3) выявление моделей управления в синтагмах; 4) упорядочение выявленных отношений по силе связи между единицами; 5) выявление синтаксически эквивалентных единиц; 6) упаковка символьных цепочек в деревья. Возможность предварительной подготовки текста позволяет решать эти задачи на разных грамматических уровнях. Интерпретация выявленных грамматических единиц и отношений осуществляется человеком.

Некоторые результаты эксперимента по решению задачи первого типа - выявление грамматики G_T для словоизменительных аффиксов русского языка - были приведены в [1]. Проиллюстрируем решение задач 2-5 на тексте T_1, представляющем собой покрытие первого типа текста T объем около 200 словоупотреблений.

Сформируем множество W из всех максимальных подцепочек текста T_1, не содержащих ни знаков препинания, ни неперекодированных слов T. Рассмотрим цепочки кодов окончаний, число вхождений которых в элементы W не меньше двух (цепочки с меньшим числом вхождений заведомо не участвуют в системе парадигматических отношений). Построим некоторые деревья сочетаемости этих цепочек в правых контекстах единичной длины (рис.3). Окончания, расположен-

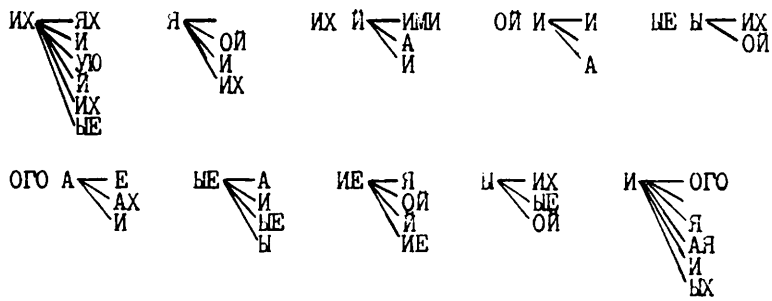

Рис. 3

ные справа от центрального соответствуют его правым контекстам. Проанализировав эти деревья с помощью ПI, выделим новые грамматические единицы - синтагмы текста T_1 (задача 2): ИХ И, ОЙ И, ЫЕ Ы, которые интерпретируются как модели управления в синтагмах текста T (задача 3). Эти модели управления могут быть упорядочены по силе связи между составляющими их элементами (задача 4) в зависимости от числа контекстов, в которых они встречаются в T_1. Взаимозамещаемые цепочки окончаний (например, ОЙ И, И; ЫЕ Ы, Ы) интерпретируются как синтаксически эквивалентные грамматические единицы (задача 5). Каждой паре таких единиц можно сопоставить некоторый инвариантный признак, свойственный обеим единицам (например, паре ОЙ И, И - родительный падеж единственного числа, паре ЫЕ Ы, Ы - винительный падеж множественного числа).

Укажем некоторые прикладные области, в которых возникают проблемы, сводимые к задачам указанных типов.

I) Дешифровка неизвестных языков и анализ текстов неязыковой природы (задачи I–5).

2) Осуществление частичного синтаксического контроля на основе использования признака согласованности произвольного заданного текста с грамматикой G_T. Правила синтаксического контроля выявляются автоматически по тексту T (задачи 2,3).

3) Обнаружение некоторых возможностей трансформации фраз текста, не нарушающих его правильности (задача 5). Например, выделение пары синтаксически эквивалентных единиц ОЙ И и И говорит о том, что в тексте T могут быть допустимы преобразования типа w_1 + ОЙ ⌴ w_2 + И ⇔ w_3 + И, где w_1 + ОЙ, w_2 + И, w_3 + И – некоторые слова текста. Условия применимости таких преобразований выявляются при помощи анализа контекстов.

Частичный синтаксический контроль и трансформацию фраз предполагается использовать в системе автоматизации редакционно-издательских работ [I3].

4) Выявление наиболее типичных парадигматических и синтагматических отношений между словами, лежащих в основе построения многих информационных языков [I4] (задачи I,2).

5) Сопоставление частотных словарей и организация словарей в памяти ЭВМ [I5] (задача 6).

6) Сравнительное изучение и формальная классификация текстов разных типов на основе введения специальных отношений на сопоставляемых тексту грамматических структурах.

7. Программы, реализующие предположения П1–П5, написаны для ЭВМ типа ЕС на языке R/TRAN. Для представления множеств W и E используются древесные структуры типа RD-деревьев [I2]. На базе построенной программной системы предполагается создание пакета прикладных программ, режимы работы которого представлены на рис.2. Общий объем программ составляет около 700 предложений R/TRANa. Время построения пространства π по дереву, состоящему из ~ II000 вершин – около I2 минут (никаких дополнительных сведений о тексте не задавалось). Время построения трех деревьев общим объемом ~ I4000 вершин производится за ~ II минут.

Литература

I. ТИМОФЕЕВА М.К. Автоматическое выявление структуры парадигматических отношений текста. –В кн.: Машинные методы обнаружения закономерностей. (Вычислительные системы, вып. 88). Новосибирск, 1981, с. I28–I35.

2. МЕЛЬНИКОВ Г.П. Системология и языковые аспекты кибернетики. - М.: Сов.радио, 1978. - 367 с.

3. МУРЯСОВ Р.З. К теории парадигматики в лингвистике.-Вопросы языкознания, 1980, № 6, с. 109-121.

4. СУХОТИН Б.В. Методы дешифровки сообщений. -В кн.: Внеземные цивилизации. М., 1969, с. 222-352.

5. РЕВЗИН И.И. Метод моделирования и типология славянских языков. -М.: Наука, 1967. - 299 с.

6. МАРКУС С. Теоретико-множественные модели языков. - М.: Наука, 1970. - 332 с.

7. ГЛАДКИЙ А.В. Формальные грамматики и языки. -М.: Наука, 1973. - 368 с.

8. АНДРЕЕВ Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. - Л.: Наука, 1967. - 403 с.

9. РЕВЗИН И.И. О некоторых вопросах дистрибутивного анализа и его дальнейшей формализации. -В кн.: Проблемы структурной лингвистики, М., 1962, с. 13-21.

10. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Методы поиска и анализ статистических закономерностей в символьных последовательностях. -В кн.: Машинные методы обнаружения закономерностей. Новосибирск, 1976, с. 75-84.

11. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. -В кн.: Вычислительные системы. Вып. 62. Ассоциативное кодирование. Новосибирск, 1975, с.49-71.

12. КОСАРЕВ Ю.Г., ЧУЖАНОВА Н.А. Автоматический синтез алгоритмов классификации словоформ по типам словоизменительных парадигм. -В кн.: Структурная и математическая лингвистика. Киев, 1978, с. 24-32.

13. КОСАРЕВ Ю.Г., МОСКВИТИН А.А. Система широкого применения для автоматизации редакционно-издательских работ. - В кн.: Методы обработки информации (Вычислительные системы, вып. 74). Новосибирск, 1978, с. 3-20.

14. МОСКОВИЧ В.А. Информационные языки. - М.: Наука, 1971. - 144 с.

15. ТИМОФЕЕВА М.К. Применение R-технологии программирования для организации больших словарей в памяти ЭВМ. -В кн.: Автоматизированные системы управления ВУЗом. Новосибирск, 1978, с. 57-66.

Поступила в ред.-изд.отд.
20 августа 1981 года