

Д.С. Бибииков

## Итеративный способ обработки текстов естественного языка с целью извлечения знаний

Описан итеративный способ как один из подходов к анализу естественно-языкового текста. Способ основан на использовании локального словаря, создаваемого из анализируемого текста, толкового словаря, информации о месте вхождения слов в предложения текста и частотных характеристик слов в нем. Извлеченные знания представляются в виде формул логики предикатов, модальной и временной логики.

Iterative method is described as one of the approaches to the analysis of natural language texts. This method is based on the use of local vocabulary, created from the sample text, a glossary, information on the occurrence of the word in a sentence and the frequency characteristics of words in this text. Extracted knowledge represented as formulas of predicate logic, modal and temporal logic.

Описано ітеративний спосіб як один з підходів до аналізу природно-мовного тексту. Спосіб ґрунтується на використанні локального словника, створюваного з аналізованого тексту, тлумачного словника, інформації про місце входження слів в пропозиції тексту і частотні характеристики слів у ньому. Добуті знання подаються у вигляді формул логіки предикатів, модальної та тимчасової логіки.

**Введение.** Проблемы, связанные с анализом естественно-языковых объектов (ЕЯО), традиционно относят к области искусственного интеллекта. Однако объективные трудности, возникающие при анализе ЕЯО, не позволяют положительно решать проблему автоматизации такого анализа. Трудности связаны с тем, что проблема анализа ЕЯО относится к областям, плохо поддающимся формализации, хотя по этой проблеме существует литература с описанием различных методов и подходов к решению частных случаев [1–5]. В работах [5, 6, 10] описывались способы извлечения первичных знаний из естественно-языковых текстов (ЕЯТ), ограничиваясь окрестностью одного предложения (именно в этом смысле употреблялось слово «первичных»).

В статье решается задача семантического анализа предложений естественного языка, не ограничиваясь рамками одного предложения. Используя информацию, которую дает система частотного анализа слов в ЕЯТ [10], итеративно выполняется семантический анализ. На первом шаге итерации все глаголы и предикатные слова объявляются отношениями, на втором шаге уточняется арность этих отношений и выполняется их семантический анализ в пределах одного предложения, на третьем шаге выполняется уточнение семантического смысла полученной

информации, исходя из некоторой локальной части текста или из текста в целом. Этот подход можно выполнять как в автоматизированном режиме, так и в диалоговом. Извлеченные из текста отношения трактуются как знания, представляются в виде формул логики предикатов, модальной пропозициональной логики и линейной временной логики. Эти формулы заносятся в базу знаний, с помощью которой проводится структурирование отношений и дальнейший (логический) анализ полученной информации. Ведущей парадигмой структурирования информации сегодня являются онтологии или иерархические концептуальные структуры, представляющие собой модель предметной области, состоящей из иерархии понятий (концептов) предметной области, связей между ними и законов, действующих в рамках этой модели. Поэтому данный анализ ЕЯТ ориентирован прежде всего на построение онтологий.

### Формальная постановка проблемы

Процесс автоматизации какой-либо деятельности, как правило, требует формализованной постановки задачи и ее анализа с целью выработки метода решения. Когда речь идет об автоматизации процесса извлечения знаний из ЕЯТ и построения соответствующей онтологии, то необходимо определить понятия «знание» и «извлечение знания». С целью формализации

вышеупомянутых понятий, введем следующие определения, пользуясь нотацией констрейнтового программирования [3].

Пусть дано некоторое множество  $D$ , на котором определена конечная совокупность  $R = \{R_1, \dots, R_k\}$  отношений  $R_i \subseteq D^n$ ,  $i = 1, 2, \dots, k$ , конечной арности. Языком ограничений  $L$  на  $D$  называется непустое множество  $L \subseteq R$ . Проблема выполнимости ограничений из  $L$  формулируется следующим образом.

Для произвольного множества  $D$  и языка ограничений  $L$  на  $D$  проблемой выполнимости ограничений  $CSP(L)$  есть решение такой комбинаторной задачи:

**дана** тройка  $P = (V, D, C)$ , где

- $V = \{v_1, v_2, \dots, v_m\}$  – конечное множество переменных;

- $C = \{c_1, \dots, c_q\}$  – конечное множество ограничений, где ограничение  $c_i$  из  $C$  является парой  $(s_i, R_i)$ , где  $s_i = (v_{i1}, \dots, v_{ij})$  – кортеж, состоящий из переменных,  $R_i \in L$  –  $n_j$ -арное отношение на  $D$ ;

**найти** функцию  $\varphi: V \rightarrow D$  такую, что  $\forall (s_i, R_i) \in C$  кортеж  $(\varphi(v_{i1}), \dots, \varphi(v_{ij})) \in R_i$  либо убедиться в том, что ее не существует,  $i = 1, 2, \dots, q$ . Множество  $D$  в этом случае называется областью проблемы, а функция  $\varphi$  – интерпретацией  $CSP(L)$ .

В случае анализа ЕЯТ, с целью извлечения знаний, множество  $D$ , как область проблемы, интерпретируется как множество объектов, извлеченных из входного текста  $T$ , которое факторизовано по некоторому отношению эквивалентности  $R$  (назовем его отношением синонимии) и в котором «закодированы» отношения  $R_i$ ,  $i = 1, 2, \dots, k$ . Множество переменных  $V = \{v_1, v_2, \dots, v_m\}$  принимает значения в этом факторизованном множестве объектов, фигурирующих в тексте  $T$  (это могут быть лексико-грамматические разряды, конкретные объекты (люди, даты, предметы и т.п.)).

**Проблемой извлечения знаний из ЕЯТ** называется проблема поиска интерпретации  $\varphi: V \rightarrow D$  с явным построением отношений  $R_i$  совокупности  $L \subseteq R$ . При этом, отношения  $R_i \in L$ ,  $i = 1, 2, \dots, k$ , извлеченные из текста  $T$ , назовем **знаниями**.

Приведенное определение достаточно общее и его необходимо уточнять для настройки на конкретную область применения. Конкретизация интерпретации и отношений в таком случае определяется целями, преследуемыми при анализе данного текста  $T$ . Хорошо известными примерами такого уточнения являются:

- **Лексико-грамматический анализ** приводит к конкретизации интерпретации  $\varphi: V \rightarrow T$  и отношений  $R_i \in L$ . Интерпретация  $\varphi$  в данном случае представляется в виде суперпозиции двух функций  $\varphi_1$  и  $\varphi_2$ , т.е.  $\varphi(V) = \varphi_2(\varphi_1(V)) = \varphi_1 * \varphi_2(V)$ , где  $*$  означает суперпозицию функций. Функции  $\varphi_1$  и  $\varphi_2$  реализуют процесс синтаксического и семантического анализа предложений текста  $T$ , а отношения  $R_1$  и  $R_2$  – это синтаксические (синтаксические правила языка, в которых представлен текст  $T$ ) и семантические ограничения.

Функцию  $\varphi_1$  тоже можно рассматривать как суперпозицию отображений  $\varphi_{1_1}$  и  $\varphi_{1_2}$ , реализующих соответственно морфологический и синтаксический анализ предложений ЕЯТ  $T$ , и которые вместе с отображением  $\varphi_2$  составляют классическую систему лексико-грамматического анализа [4].

- **Силлогистика Аристотеля** – другой пример уточнения интерпретации  $\varphi$  и отношений  $R_i \in L$ . В этом случае интерпретация  $\varphi$  носит теоретико-множественный характер, а отношения  $R_i \in L$  – это отношение включения для множеств и его свойства. Более полное описание этого уточнения можно найти в работах [5–7].

- **Текст библиографического характера** – пример хорошо структурированного текста. Это значит, что проблему извлечения знаний из такого текста можно решить в автоматизированном режиме.

### **Пример автоматизированной обработки ЕЯТ**

Рассмотрим пример автоматизированной обработки фрагмента ЕЯТ. Подсистема частотного анализа использует толковый словарь  $S$  естественного языка  $L(X)$  (это может быть словарь русского, украинского, английского или какого-либо другого естественного языка).

Текст  $T$  состоит из предложений языка  $L(X)$  и представляет текст, не содержащий никаких символов, кроме символов алфавита  $X$  (т.е.  $T$  не содержит формул, графиков, рисунков и т.п.). Эта подсистема реализует отношение  $\gamma$ , состоящее из суперпозиции двух отношений  $\gamma_1 * \gamma_2$ , выполняемых последовательно. Содержательно отношение  $\gamma_1$  означает распознавание принадлежности слова к данному языку и проверку правильности написания слова  $t_{ij} \in t_i$ , где  $t_i \in T$ , в соответствии с написанием его в толковом словаре, т.е.

$$\gamma_1(t_{ij}) = \begin{cases} 1, & \text{если } t_{ij} \in S; \\ 0, & \text{если } t_{ij} \notin S. \end{cases}$$

Если слово  $t_{ij} \in t_i$  распознано в словаре  $S$ , то оно заносится в словарь  $T'$  правильных слов, а если это не так, то предусматривается сигнализация о том, что данное слово отсутствует в словаре  $S$  и принимается решение о добавлении данного слова в словарь или его исправлении (слово может быть искажено, например, вследствие сканирования текста  $T$ ).

Словари  $S$  и  $T'$  – входные данные для отношения  $\gamma_2$ . Содержательный смысл отношения  $\gamma_2$  сводится к тому, что если  $\gamma_1(t_{ij}) = 1$ , то  $\gamma_2(t_{ij})$  определяет его грамматическую единицу языка (имя собственное, сказуемое, существительное, числительное и т.п.) а также возможные флексии слова  $t_{ij} \in t_i$ .

Областью интерпретации текста  $T$  есть модель  $A = (D, \Pi)$ , где  $T$  – это исходный текст, возможно расширенный некоторой дополнительной информацией, а сигнатура предикатов (отношений)  $\Pi$  определяется из текста  $T$  в результате использования информации о различных вхождениях слова  $t_{ij}$  в предложения  $t_i \in T$ . При этом вычисление отношения  $\phi$  ограничивается отдельно взятым предложением  $t_i \in T$ , определяемым каждым вхождением слова  $t_{ij}$  в текст  $T$ . В случае трудности определения предиката  $\pi_i \in \Pi$ , предусматривается диалоговый режим вычисления  $\phi(\pi_i)$  и  $\gamma(\phi(\pi_i))$ .

Предлагаемая система анализа схематически выглядит так, как показано на рис. 1.

В этой схеме отношение  $\gamma = \gamma_1 * \gamma_2$  вычисляет программа  $P_1$ . Результат ее работы – два файла  $F_1$  и  $F_2$ , заполненные соответственно числовыми характеристиками слов входного текста  $T$  и словами  $t_{ij}$  предложений этого текста. Структура файла  $F_2$  показана ниже на рис. 2.

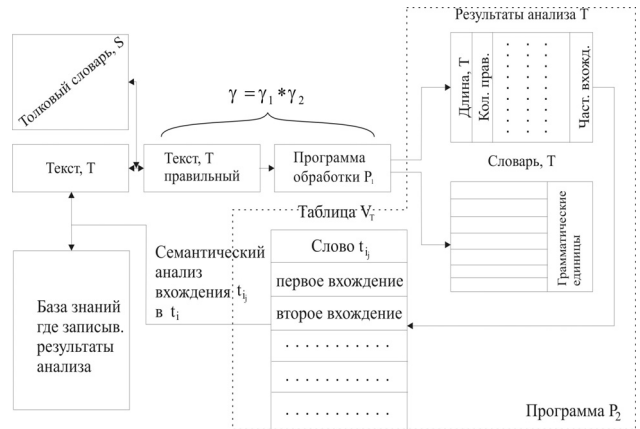


Рис. 1. Схема системы анализа предложений естественного языка

Слова	Длина	Частота	Часть речи
зовнішньополітична	18	1	прикметник
використовуватися	17	1	дієслово
загальнознаними	17	1	прикметник
континентального	16	1	прикметник
багатоманітності	16	1	іменник
недоторканність	15	1	іменник
конституційного	15	8	прикметник
недоторканності	15	1	іменник
взаємовідного	15	1	прикметник
спіробітництва	15	1	іменник
самоврядування	14	4	іменник
конституційний	14	1	прикметник
конституційних	14	1	прикметник
обов'язковість	14	1	іменник
функціонування	14	2	іменник
господарювання	14	1	іменник
чорнобильської	14	1	прикметник

Рис. 2. Структура файла  $F_2$

Файлы  $F_1$  и  $F_2$ , сформированные программой  $P_1$ , служат входными данными для работы программы  $P_2$ , вычисляющей отношение  $\phi$ . При этом работа программы  $P_2$  сводится к построению таблицы  $V_T$  для слов  $t_{ij} \in t_i$ ,  $t_i \in T$ . Затем по этой таблице и предложениям текста  $T$  определяется семантический смысл рассматриваемого предложения. Предложение  $t_i \in T$  оп-

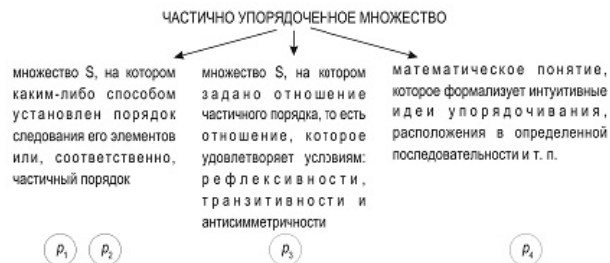
ределяется на основе номера вхождения слова  $t_{ij}$  в текст  $T$  с помощью таблицы  $V_T$ , вид которой приведен на рис. 3.

Слова	№ вхождения	№ предложени
національно-культурних	1	57
нормативно-правові	1	39
зовнішньополітична	1	86
використовуватися	1	63
загальновізнаними	1	86
континентального	1	59
багатоманітності	1	71
недоторканність	1	9
конституційного	1	21
конституційного	2	25
конституційного	3	29
конституційного	4	52
конституційного	5	81
конституційного	6	93
конституційного	7	95
конституційного	8	96
недоторканності	1	79

Рис. 3. Структура таблицы  $V_T$

### Обработка определений в ЕЯТ и их формальное представление

При анализе ЕЯТ первоочередным является обнаружение двух фундаментальных отношений, присутствующих практически в любом ЕЯТ. Это отношения эквивалентности и частичного порядка. Первое из этих отношений определяет классы синонимичных объектов, а второе – иерархию подчиненности классов эквивалентности. Оба отношения составляют основу построения онтологий, а знания, полученные на этом этапе, назовем первичными. В отношении частичного порядка может вкладываться различный семантический смысл: это может быть *отношение таксономии* (принадлежать множеству, классу, группе и т.п.), *отношение партномии* (состоит из), *отношение генеалогии* (отец – сын), *причинно-следственное отношение* (если – то), *атрибутное отношение* и др. [9]. В качестве примера рассмотрим множество математических определений. Рассмотрим, взятые в Большой советской энциклопедии, а также в ряде специализированных учебников такие определения: частично упорядоченное множество, линейно-упорядоченное множество, вполне упорядоченное множество.



Здесь:

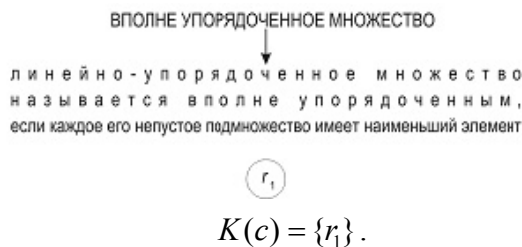
$p_1$  = «множество  $S$ , на котором каким-либо способом установлен порядок следования его элементов»,  $p_2$  = «множество  $S$ , на котором каким-либо способом установлен частичный порядок»,  $p_3$  = «множество  $S$ , на котором задано отношение частичного порядка, т.е. отношение, которое удовлетворяет условиям: рефлексивности, транзитивности и антисимметричности»,  $p_4$  = «математическое понятие, формализующее интуитивные идеи упорядочивания, расположения в определенной последовательности и т.п.».

$K(a) = \{p_1, p_2, p_3, p_4\}$  – класс эквивалентности.



$$K(b) = \{q_1, q_2, q_3\},$$

где:  $q_1$  = «частично упорядоченное множество  $S$ , в котором для любых двух элементов  $a$  и  $b$  имеет место  $a \leq b$ »,  $q_2$  = «частично упорядоченное множество  $S$ , в котором для любых двух элементов  $a$  и  $b$  имеет место  $b \leq a$ »,  $q_3$  = «множество  $S$  называется линейно-упорядоченным, если на нем задано некоторое отношение линейного порядка».



Здесь  $r_1$  полностью совпадает с определением:  $r_1$  = «линейно-упорядоченное множество на-

зывается вполне упорядоченным, если каждое его непустое подмножество имеет наименьший элемент».

Приведенные примеры показывают, что построение классов эквивалентности не составляет особых трудностей. В результате построения классов появляются объекты:

$$K(a) = \{p_1, p_2, p_3, p_4\}, K(b) = \{q_1, q_2, q_3\}, \\ K(c) = \{r_1\}.$$

Проблема появляется при вычислении второго отношения, определяющего отношение подчиненности (иерархии) между имеющимися классами эквивалентности. Однако при таком абстрактном представлении классов  $K_i(x)$  это отношение определить нельзя. Для этого необходимо знать структурные характеристики элементов из классов  $K_i(x)$ . Поэтому естественным образом появляется необходимость в структуре элементов из классов эквивалентности. Например, если вернуться к вышерассмотренным примерам, то каждый элемент из класса  $K(a)$  принимает вид:

$$p_1 = (p_{1_1}, p_{1_2}, p_{1_3}), p_2 = (p_{2_1}, p_{2_2}),$$

$$p_3 = (p_{3_1}, p_{3_2}, p_{3_3}, p_{3_4}, p_{3_5}), p_4 = (p_{4_1}, p_{4_2}, p_{4_3}),$$

где  $p_{1_1}$  = «множество»,  $p_{1_2}$  = «установлен порядок следования элементов»,  $p_{1_3}$  = «установлен частичный порядок»,  $p_{2_1}$  = «множество»,  $p_{2_2}$  = «установлен частичный порядок»,  $p_{3_1}$  = «множество»,  $p_{3_2}$  = «задано отношение частичного порядка»,  $p_{3_3}$  = «удовлетворяет рефлексивности»,  $p_{3_4}$  = «удовлетворяет транзитивности»,  $p_{3_5}$  = «удовлетворяет антисимметричности»,  $p_{4_1}$  = «математическое понятие»,  $p_{4_2}$  = «формализует интуитивные идеи упорядочивания»,  $p_{4_3}$  = «формализует интуитивные идеи расположения».

Аналогично структурируются и остальные элементы в классах эквивалентности.

$$q_1 = (q_{1_1}, q_{1_2}, q_{1_3}, \dots), q_2 = (q_{2_1}, q_{2_2}, q_{2_3}, \dots), \\ q_3 = (q_{3_1}, q_{3_2}, q_{3_3}, \dots), r_1 = (r_{1_1}, r_{1_2}, r_{1_3}, \dots).$$

Из приведенной структуризации вытекает следующая формализация. Если класс эквивалентности относится к объекту  $a$ , то его формальное определение выглядит как дизъюнкция элементов, составляющих этот класс. Каждый элемент, входящий в тот или иной класс эквивалентности описывается соответствующим предикатом, т.е. если  $K(a) = \{p_1, p_2, p_3, p_4\}$ , то  $p(a) \Leftrightarrow p_1(a) \vee \dots \vee p_4(a)$ , где  $p_i$ -предикаты, характеризующие элементы из класса  $K(a)$ , а их дизъюнкция характеризует весь класс понятия  $a$ .

Далее, если  $q_i \in K(a)$  и  $q_i = (q_{i1}, q_{i2}, \dots, q_{ik})$ , то элемент  $p_i$  (или объект  $p_i$ ), характеризующийся атрибутами  $p_{ij}$ , представляется в виде конъюнкции

$$p_i(a) \Leftrightarrow p_{i1}(a) \wedge \dots \wedge p_{ik}(a),$$

где  $p_{ij}(a)$  – предикат, характеризующий отдельный атрибут понятия  $a$ ,  $i = 1, \dots, l; j = 1, \dots, k$ .

Следовательно, каждый класс  $K(a)$  описывается дизъюнктивной формой вида

$$p(a) \Leftrightarrow (p_{11}(a) \wedge \dots \wedge p_{1m_1}(a)) \vee \dots \\ \vee (p_{l1}(a) \wedge \dots \wedge p_{lm_l}(a)).$$

Введенная формализация определяет отношение частичного порядка, которое вводится следующим образом:

$$K(a) \leq K(b) \Leftrightarrow (\exists p_i(a))(\exists q_j(b))(q_j(b) \leq p_i(a)),$$

где  $q_j(b) \leq p_i(a)$  означает, что  $q_j(b)$  входит в виде конъюнктивного члена в  $p_i(a)$ .

Введенные таким образом отношения частичного порядка естественным образом требуют предикатно-реляционного представления объектов из классов эквивалентности и самих этих классов [8]. Иллюстрируя сказанное, вернемся к вышеприведенному примеру. Выясним, как определяется тот факт, что классу «частично упорядоченное множество» подчиняется класс «линейно-упорядоченное множество».

Класс «частично упорядоченное множество», обозначенный как  $K(a)$ , описывается формулой

$$p(a) \Leftrightarrow p_1(a) \vee p_2(a) \vee p_3(a) \vee p_4(a),$$

$$\text{где } p_1(a) \Leftrightarrow \text{МНОЖ-УСТАН-ЧАСТ-ПОР}(a),$$

$$p_2(a) \Leftrightarrow \text{УСТАН-ПОР-СЛЕДОВ}(a), p_3(a) \Leftrightarrow$$

$$\text{УДОВЛ-РЕФЛ-АНТИСЕМ-ТРАНЗИТ}(a),$$

$$p_4(a) \Leftrightarrow \text{ФОРМ-ИНТ-ИД-УПОРЯД}(a),$$

Класс «линейно-упорядоченное множество», обозначенный  $K(b)$ , описывается формулой

$$q(b) \Leftrightarrow q_1(b) \vee q_2(b) \vee q_3(b) \vee \dots \vee q_l(b),$$

где  $q_1(b) \Leftrightarrow p_1(a) \wedge \text{УСТ-ЛИН-ПОРЯД}(b) \wedge \text{НЕ-РАВЕН-А-МЕНЬШ-В}(b) \wedge \dots$

Используя определенное выше отношение частичного порядка, находим, что  $p_1(a)$  входит в определение класса  $K(a)$ , а это значит, что  $K(b) \leq K(a)$ .

Аналогично определяется подчинение класса  $K(c)$  классу  $K(b)$ , в результате чего получаем граф  $K(a) \rightarrow K(b) \rightarrow K(c)$ .

Полученную таким образом иерархию можно изменять или модифицировать путем диалога с пользователем с целью достижения более правильного представления.

Для полноты картины рассмотрим понятия «Отношения квазипорядка» и «Отношения строгого порядка» и определим типы связей понятий.

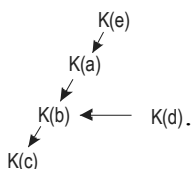
Отношение квазипорядка – это бинарное отношение на множестве, обладающее свойствами рефлексивности и транзитивности.

Аналогично описанному методу получаем новый класс понятия  $K(d)$ , и, соответственно, его структуру  $K(d) = \{s_1\}$ , где  $s_1 = (s_{1_1}, s_{1_2}, s_{1_3})$ .  $s_{1_i}$  соответственно равны: «бинарное отношение», «обладает свойством рефлексивности», «обладает свойством транзитивности».

Отношение строгого порядка – это бинарное отношение на множестве, обладающее свойствами транзитивности и иррефлексивности.

Здесь имеем дело уже с классом  $K(e)$ , и соответственно его структурой  $K(e) = \{t_1\}$ .  $t_1 = (t_{1_1}, t_{1_2}, t_{1_3})$ , где:  $t_{1_1}$  – «бинарное отношение»,  $t_{1_2}$  – «обладает свойством иррефлексивности»,  $t_{1_3}$  – «обладает свойством транзитивности».

Анализируя введенные новые понятия получаем новый граф вида:



С классами  $K(e)$ ,  $K(a)$ ,  $K(b)$ ,  $K(c)$  связано понятие сильной связи, поскольку присуща вложенность свойств, а классам  $K(b)$ ,  $K(d)$  применимо понятие слабой связи, поскольку строгой вложенности нет, а пересечение свойств дает непустое множество.

Итак, вышеописанный процесс обработки текстов определений формализуется следующим образом.

Пусть  $T$  – множество текстов определений. По этому множеству  $T$  строится множество классов, которое определяется отношением эквивалентности  $R$  и составляет фактор множество  $D$ . На полученном так множестве  $D$  определяются специальные отношения  $R_2, \dots, R_k$ , описывающие характеристические свойства элементов из  $D$ , т.е. элементов из классов эквивалентности. Эти отношения представляются в виде предикатов, которые определяют отношение частичного порядка  $R_1$ . Это отношение – второе отношение, построения онтологии. Более точно, онтология строится по транзитивному замыканию  $R_1^*$  отношения  $R_1$ , согласованного с отношением  $R$ .

**Определение 1.** Отношение  $R$  и  $R_1^*$  назовем согласованным, если  $\forall a, b \in D$  имеет место включение  $(a, b) \in R * R_1^*$ , где  $R * R_1^*$  – суперпозиция отношений  $R$  и  $R_1^*$ .

Из этого определения естественным образом следует первичная онтология:  $O = (D = T/R, \mathfrak{R} = \{R_1, R_2, \dots, R_k\}, \varphi, A)$ , где  $\varphi: D \rightarrow T$  – интерпретация,  $A$  – множество аксиом, определяемое предикатами, описывающими характеристические свойства элементов из  $D$ ,  $R_2, \dots, R_k$  – соответствующие им отношения, а  $R_1$  – отношение частичного порядка.

### Обработка ЕЯТ относительно временных и объектных отношений

Рассмотрим еще один пример конкретизации отношений  $R$  и  $R_1$ , которые назовем временными и объектными.

Отношение эквивалентности  $R$  определяется конкретным объектом (отсюда и название этого отношения), фигурирующим в тексте  $T$  (на-

пример, личность, театр, институт, вуз, кафедра и т.д.), а все объекты, связанные с ним в тексте  $T$ , составляют класс эквивалентности по этому отношению. Отношение  $R_1$  определяет связь объектов из классов эквивалентности отношения  $R$  либо в хронологическом порядке (временные зависимости), либо генеалогическом, либо каком-нибудь ином подобном порядке.

Отношение  $R$  и  $R_1$  – источник построения новых отношений. Например, если  $R_1$  описывает временные зависимости, то можно определить отношение, связывающие объекты, относящиеся к данному конкретному времени (год, месяц, день и т.п.). Поясним это примером.

Пусть  $D = \{a, b, c, d\}$  – объекты, фигурирующие в данном тексте, который обрабатывается. Тогда  $D/R = \{K(a), K(b), K(c), K(d)\}$  состоит из классов, элементами которых есть моменты времени  $t_{xi}$ , связанные с объектом  $x \in D$ ,  $i = 1, \dots, j_x$ . Допустим, что нас интересуют объекты, фигурирующие в моменты времени  $t$ . Тогда момент времени  $t$  определяет отношение  $R_t$ , состоящее из объектов, связанных этим моментом времени. Арность этого отношения определяется мощностью множества  $D$ . Если  $D = \{K(a), \dots, K(b)\}$ , то  $R_t = \{(a', \dots, b') \mid a' \in K(a) \wedge \dots \wedge b' \in K(b)\}$ .

На отношениях  $R_t$  определяются отношения линейного порядка:  $R_t \leq R_{t'} \Leftrightarrow t \leq t'$ . Пользуясь этим отношением порядка, можно ввести следующее понятие.

**Определение 2.** *Элементарным временным сценарием* для временного интервала  $[t_1, t_k]$  называется цепь  $R_{t_1} \leq R_{t_2} \leq \dots \leq R_{t_k}$ , где  $R_{t_i}$  – отношения, определенные ранее.

Это определение можно модифицировать в зависимости от семантики отношений  $R$  и  $R_1$ . Действительно, цепь отношений  $R_{t_1} \leq R_{t_2} \leq \dots \leq R_{t_k}$  описывает хронологию некоторых событий во времени, связанных с конкретными людьми, объектами и т.д. Такого типа цепь в действительности может служить основой некоторого реального сценария.

**Заключение.** Описанные в статье способы автоматизации обработки ЕЯТ составляют основу как теоретического, так и практического анализа извлечения знаний. Используя эту основу и прежде всего ее реализацию, предполагается наращивание ее мощности за счет построения новых мета отношений над построенными отношениями, являющимися отдельными частями знаний, имеющихся в исследуемом тексте.

1. Палагин А.В., Кривый С.Л., Петренко Н.Г. Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация // УСиМ. – 2009. – № 3. – С. 42–55.
2. Палагин А.В., Петренко Н.Г. Системно-онтологический анализ предметной области // Там же. – 2009. – № 4. – С. 3–14.
3. Cohen D., Jeavons P. The Complexity of Constraint Languages // Handbook of Constraint Programming. – Eds. by F. Rossi et al. – 2006. – P. 245–280.
4. Апресян Ю.Д. Лингвистический процессор для сложных информационных систем. – М.: Наука. – 1992. – 324 с.
5. Алгебро-логічний підхід до аналізу та обробки текстової інформації / О.В. Палагін, С.Л. Кривий, М.Г. Петренко та ін. // Проблеми програмування. – 2010. – № 2–3. – С. 318–329.
6. Формально-логічний підхід до побудови системи аналізу знань в різних предметних областях / О.В. Палагін, С.Л. Кривий, М.Г. Петренко та ін. // Там же. – 2010. – № 2–3. – С. 382–389.
7. Кулик Б.А. Логика естественных рассуждений. – С.-Петербург: Невский диалект. – 2001. – 127 с.
8. Рубашкин В.Ш. Представление и анализ смысла в информационных системах. – М.: Наука. – 1989. – 188 с.
9. Gavrilova T., Laird D. Practical Design of Business Enterprise Ontologies / Indust. Appl. of Semantic Web. – Eds. by M. Bramer, V. Terzyn. – Springer, 2005. – P. 61–81.
10. Палагин А.В., Кривый С.Л., Бибииков Д.С. Обработка предложений естественного языка с использованием словарей и частоты появления слов // Natural and Artificial Intel. Intern Book Series. – Intel. Proc. – ITNEA. – Sofia. – N 9. – 2010. – P. 44–52.

Поступила 25.07.2011  
Тел. для справок: (044) 526-6422 (Киев)  
E-mail: gbbcoff@gmail.com  
© Д.С. Бибииков, 2012