Модели знаний и данных в координатном представлении

Емельяненков В.И.
Кафедра телекоммуникаций и информационных технологий
Белорусский государственный университет
Минск, Республика Беларусь
emelvi@bsu.by

***Аннотация*—В данной работе рассматриваются вопросы трехмерного координатного представления онтологических моделей, схемы концептов которых могут быть заданы как ациклические иерархические структуры. Используются механизмы и методы отображения узлов решетки на структуру растра, которые позволяют использовать координаты точек растра в качестве индексов концептов задаваемых моделей данных.**

Ключевые слова: иерархические структуры, онтологические модели данных, решетки, индексирование.

В задачах обработки данных распространение получили онтологии, представимые с помощью дескрипционных логик. Рассмотрим логику $\mathcal{ALC}$ с ацикличной терминологией [1].

Определение 1. (Синтаксис). Множество концептов логики $\mathcal{ALC}$ задается следующим определением:

- символы $\top$ и $\bot$ концепты (истина и ложь);
- всякий атомарный концепт A, B, C, … является концептом;
- если C концепт, то $\neg C$ концепт;
- если C и D концепты, то $C \sqcap D$ и $C \sqcup D$ концепты;
- если C концепт, а R атомарная роль, то $\exists R.C$ и $\forall R.C$ концепты;
- никакие другие выражения не являются концептами.

Определение 2. (Семантика). Интерпретация есть пара $I = (\Delta, {}^I)$, состоящая из непустого множества Δ, называемого областью данной интерпретации, и интерпретирующей функции I, которая сопоставляет:

- каждому атомарному концепту $A \in \mathrm{CN}$ произвольное подмножество $A^I \subseteq \Delta$;
- каждой атомарной роли $R \in \mathrm{RN}$ произвольное подмножество $R^I \subseteq \Delta \times \Delta$.

В данной работе в качестве интерпретации исследуется растровое представление булевой решетки, как это изображено на Рис 1 [2]. В данном случае оно используется как эквивалент теоретико-множественной интерпретации концептов логики $\mathcal{ALC}$. Поскольку это изоморфный эквивалент булевой решетки, в нем выполняются необходимые свойства:

- $\top^I = \Delta$, $\quad \bot^I = \varnothing$, $\quad (\neg C)^I = \Delta \backslash C^I$;
- $(C \sqcap D)^I = C^I \cap D^I$, $\quad (C \sqcup D)^I = C^I \cup D^I$;
- $(\exists R.C)^I = \{e \in \Delta \mid$ существует $d \in \Delta$ такой, что $\langle e, d \rangle \in R^I$ и $d \in C^I\}$;
- $(\forall R.C)^I = \{e \in \Delta \mid$ для всех $d \in \Delta$ таких, что $\langle e, d \rangle \in R^I$, выполнено $d \in C^I\}$.

Проиллюстрируем это на примерах.

Примем в качестве базиса следующий набор атомарных концептов A, B, C, D, E.

Корневой узел с координатой (0, 1) определим как пустое множество $\varnothing$, что отвечает интерпретации $\bot^I = \varnothing$. Область интерпретации $\top^I = \Delta$ в данном случае не отображена, поскольку она расположена за пределами рисунка. Далее по дуге, выходящей из узла (0, 1) следуют узлы a, b, c, d, e. Они и являются интерпретациями атомарных концептов A, B, C, D, E. В частности, это узлы $\varnothing \subseteq a$, $\varnothing \subseteq b$, $\varnothing \subseteq c$, $\varnothing \subseteq d$, $\varnothing \subseteq e$ порождающего множества. Для любой пары этого набора имеют место равенства $a \cap b = \varnothing$, $b \cap c = \varnothing$, $b \cap d = \varnothing$, $c \cap d = \varnothing$ и т.д. Затем в каждой из дуг, идущей из узлов a, b, c, d, e, содержатся узлы, помеченные двойными символами ab, ae, ad, bd, bc и ce. Эти узлы интерпретируются как объединения атомарных концептов следующим образом: ab есть $(A \sqcup B)^I = a \cup b = A^I \cup B^I$. Аналогичные равенства имеют место и для остальных узлов ae, ad, bd, bc и ce.

Поскольку рассматриваемый растр представляет дерево, то для двойных символов в нем не будет доставать одной дуги. Этот недостаток восполняется тем, что спецификация этих дуг ведется отдельно. В качестве примера в нижней части рисунка они помечены как пары точек. В частности, имеются следующие такие пары ab, ae, ad и ce. В принципе, это можно и не делать, поскольку добавляемый символ сам по себе образует гиперссылку.

Легко заметить, что для каждой пары рассмотренных узлов может быть задана операция, которая будет рассматриваться как пересечение множеств, удовлетворяющая приведенным выше правилам интерпретации.

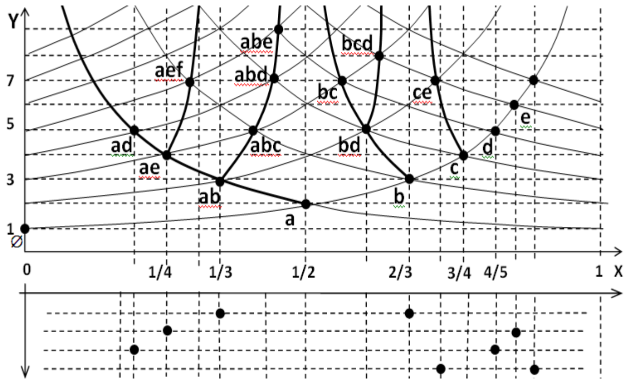

Рис. 1. Отображение концептов на узлы растра.

Например, пересечение для узлов ab и ad будет записано следующим образом: $(a \cup b) \cap (a \cup d) = a$.

Нетрудно проверить, что это является интерпретацией концептов ($A \sqcup B$) ⊓ ($A \sqcup D$)= A.

Аналогично для узлов *ad* и *bd* получим узел d в качестве их пересечения по правилам решетки. Узел d является предшественником как для *ad* так и для *bd*. Но, здесь надо принять во внимание, соединение *d* с *ad* осуществляется по гиперссылке в нижней части рисунка.

Далее аналогичным образом рассматривается процесс формирования узлов с тройными символами. Эти узлы расположены на дугах, выходящих из узлов с двойными символами. В конечном итоге по индукции за конечное число шагов формируется решетка, которая является интерпретацией для всего множества концептов в пределах заданной терминологии.

Таким образом, данное растровое представление является достаточным для задания схемы TBox логики $\mathcal{ALC}$.

Для того чтобы двигаться дальше, мы используем тот факт, что координаты $X_{i/r}$ узлов на оси X являются элементами множества рациональных чисел из интервала (0, 1). На рис. 1 в качестве примера приведены координаты 1/4, 1/3, 1/2, 2/3, 3/4 и 4/5, соответственно, для узлов *ae*, *ab*, *a*, *b*, *c* и *d*.

В работе [2] рассмотрен алгоритм восстановления любого узла в растровой решетке по его координате. Рассмотрим этот механизм в общих чертах.

Имеет место равенство $x_q = i/r$, где i – номер кривой Y^0_i, а r – номер линии равномерной шкалы. Поэтому индекс в обозначении $Y_{i/r}$ можно рассматривать как запись значения координаты $x_q = i/r$ в виде структуры i/r.

На это множество $\{Y_{i/r}\}$ и отображаются узлы иерархической структуры.

Идентификация концепта осуществляется по его координате $x^*_q = i^*/r^*$. По построению координата любого узла лежит внутри определенного интервала, границами которого являются ближайшие слева и справа к i^*/r^* узлы концептов предшествующего уровня. Один из них будет родительским i^p/r^p, а второй сопряженным с ним i^s/r^s. При этом координатой родительского узла будет та из двух, у которой значение r больше ($r^p > r^s$). Далее в свою очередь координата этого родительского узла попадает в интервал, одна из границ которого определяет последующий родительский узел. И так можно продолжать вплоть до узла концепта первого уровня.

Также возможен и вариант определения родительского узла путем прямого вычисления его координаты по заданным значениям $x^*_q = i^*/r^*$.

Благодаря этому растровая конфигурация TBox может быть представлена как множество точек координатной оси. А это уже позволяет построить координатное представление онтологии в целом, так как изображено на рис 2.

Здесь представлена конфигурация с индивидами и бинарными отношениями между ними.

По оси X откладываются значения координат узлов растра, соответствующие схеме TBox проектируемой онтологии. Помещенная внизу схема растра используется для навигации и прослеживания связи с рис. 1.

Вторая координатная составляющая ID вместе с осью X образует плоскость «Индивиды», на которой представлены элементы концептуальных классов.

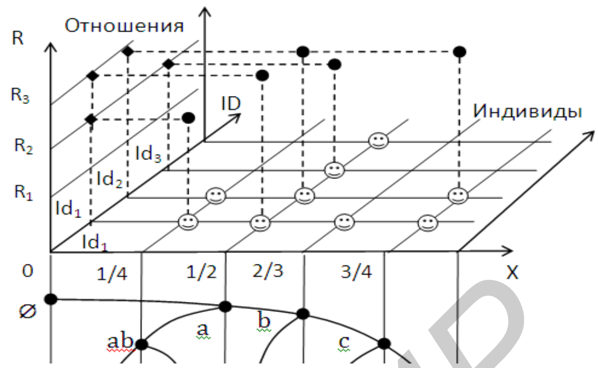

Рис. 2. Трехмерное представление онтологии.

Так индивиды концепта $A \sqcup B$, отображенного на узел растровой решетки ab помечены в плоскости «Индивиды» координатами (1/4, Id_1) и (1/4, Id_2). Соответственно, концепт A имеет четыре индивида (1/2, Id_1), (1/2, Id_2), (1/2, Id_3) и (1/2, Id_4) , концепт B имеет один индивид (2/3, Id_1) и концепт C имеет индивиды (3/4, Id_1), (3/4, Id_2).

Отношения между индивидами задаются в виде ролей на координатной составляющей R. Все индивиды, находящиеся в определенном отношении, проецируются на соответствующую плоскость. Так в отношении R_2 находятся индивиды (1/4, Id_1) и (1/2, Id_3), а в отношении R_3 обозначены три индивида (1/2, Id_1), (1/2, Id_2) и (3/4, Id_2).

Итак, получено отображение концептов схемы TBox на точки растра, образующее одномерный эквивалент терминологии, что позволяет задать линейное перечисление элементов иерархической структуры. На этой основе образуется трехмерное представление логики $\mathcal{ALC}$, включающее наборы индивидов и отношений между ними.

Как известно, одной из распространенных технологий Semantic Web являются языки OWL (Web Ontology Language), которые базируется на рассмотренной логике. Для них полученное представление позволяет достаточно эффективно визуализировать схемы прикладного анализа данных.

Кроме того в данном представлении содержатся дополнительные возможности автоматизации процедур выполнения запросов. В рассматриваемой растровой координатной системе задача выделения нужного подграфа и сопоставление его с шаблоном запроса может решаться также и с помощью рассмотренных механизмов идентификации концептов.

[1] Baader F., Calvanese D., McGuiness D.,Nardi D., Patel-Schneider *P.* The Description Logic Handbook. Theory, implementation. and applications — New York: Cambridge University Press, 2003

[2] Емельяненков В.И. Координатное представление онтологических моделей /В.И. Емельяненков //БГУИР Информационные технологии и системы (ITS'2011): Материалы Междунар. Науч. Конф. Мн. (26 октября 2011 г.). С. 111-112.