

УДК 519.7

С.Ю. Шабанов-Кушнарченко, Ю.С. Новиков

Харьковский национальный университет радиоэлектроники, Харьков

ОБОБЩЕННАЯ ПРЕДИКАТНАЯ МОДЕЛЬ ПРОЦЕССА ВЫЯВЛЕНИЯ НЕЯВНЫХ ЗНАНИЙ В ИЕРАРХИИ «ДАННЫЕ-ИНФОРМАЦИЯ-ЗНАНИЯ»

Рассмотрена проблема представления неявных зависимостей на основе трансформации графовых, а также алгебро-логических моделей явных знаний. Предложен многоуровневый подход к выявлению неявных знаний, который основывается на известной концепции трансформации данных в информацию, информации – в знания, а знаний – в метазнания. Предложенный подход обосновывает возможность получения неявных знаний на основе использования моделей явных знаний, а также структурированных массивов данных.

Ключевые слова: алгебра конечных предикатов, реляционная сеть, знания.

Введение

Извлечение и формализация скрытых зависимостей из структурированных массивов данных в задачах data-, process-, web - mining основывается на неявном влиянии таких зависимостей на сформированную структуру данных. Полученные в результате решения указанных задач структуры, например, модели процессов, представляются в виде графа и, фактически, представляют собой явные знания о процессах (либо структурированных объектах) предметной области. В то же время, модели идентичных процессов (объектов), полученные в результате решения задач интеллектуального анализа массивов данных за различные интервалы времени, во многих случаях отличаются. Такие отличия в структуре полученных графов определяются применением неформализованных, скрытых знаний при выполнении процессов и формировании исходных массивов данных. Это и указывает на принципиальную возможность выделения неявных зависимостей при анализе структурированных объектов, полученных в результате исследования массивов данных.

Основной раздел

Традиционные подходы к анализу данных и процессов основаны на известной иерархической модели DIKW «данные-информация-знания-мудрость (метазнания)» Концепция DIKW опирается в первую очередь на наблюдения за естественным интеллектом [1]. В данной работе эта модель представляет интерес в силу того, что неявные знания характерны именно для человеческого интеллекта. Преобразование неявных знаний в явную форму позволяет использовать их в системах искусственного интеллекта.

Первый уровень модели DIKW содержит данные как набор символов, которые могут быть использованы, однако сами по себе не имеют значения. Данные являются просто результатом наблю-

дений и не всегда представлены в удобном для использования формате.

$$\{p_1, p_2, \dots, p_n\}, \quad (1)$$

где $p_i, i = \overline{1, n}$ – множество данных.

На уровне информации между данными определяются связи, которые задают значение совокупности данных. На данном уровне не определяется, будет ли это значение использоваться, однако на основе информации могут быть сделаны определенные выводы.

Предикат I задает связи между данными $\{p_1, p_2, \dots, p_n\}$, чем определяет информацию об этих данных.

$$I(p_1, p_2, \dots, p_n). \quad (2)$$

Предикат I при необходимости можно представить в виде системы бинарных предикатов $I_i, i = \overline{1, m}$. Число m заранее не известно, оно определяется конкретной задачей (во-первых, самим предикатом I , во-вторых – процессом бинаризации предиката I , т.к. бинаризация практически всегда требует введения дополнительных переменных).

Накопление знаний на основе данных и информации о них можно представить формально как добавление новых предикатов, заданных на множестве исходных данных $\{p_1, p_2, \dots, p_n\}$, и создание системы

$$\{I_i, i = \overline{1, k}\} \quad (3)$$

предикатов, каждый из которых определяет информацию о данных $\{p_1, p_2, \dots, p_n\}$.

Уровень метазнаний объединяет все предыдущие, позволяя находить принципиально новое понимание из существующего знания. Формально уровень метазнаний можно представить в виде предиката второго порядка (предиката от предикатов), заданного на множестве $\{I_i, i = \overline{1, k}\}$,

$$M(I_1, I_2, \dots, I_k),$$

который связывает всю формализованную информацию о данных $\{p_1, p_2, \dots, p_n\}$ и позволяет вычислять новую информацию.

Объединяя рассмотренные ранее особенности различных форм знания, выделим его ключевые характеристики, которые позволят получить формальную модель процесса преобразования данных в явные знания с использованием неявной формы знаний. Такая формальная модель служит базой для разработки методов выделения неявных зависимостей на основе анализа информации, содержащейся в структурированных массивах данных.

К таким ключевым характеристикам знания в различных формах относятся:

- сочетание явной и неявной формы знаний, которые динамически переходят из одной в другую;
- взаимосвязь неявного знания и возможностей его практического применения;
- процессный характер неявного знания.

Предикатная модель процесса выявления неявных зависимостей, которая разработана в соответствии с данным подходом, включает в себя следующие преобразования:

- данных в информацию;
- информации в явные знания;
- данных и явных знаний в неявные знания в форме предикатов.

При первом преобразовании, данных в информацию, исходные данные представляются в виде наборов разнородных исходных элементов $p_i, q_i, r_i, i = \overline{1, K}$, которые подлежат агрегированию, уточнению, дополнению и преобразованию в информацию:

$$\{\{p_1, p_2, \dots, p_k\}, \{q_1, q_2, \dots, q_k\}, \dots, \{r_1, r_2, \dots, r_k\}\} \quad (4)$$

Преобразование осуществляется путем интерпретации на основе использования явных знаний. Последние могут иметь различные формы представления, однако в обобщенном виде их целесообразно представить в виде набора предикатов, описывающих граф преобразования данных в информацию. Графовое представление обладает тем преимуществом, что сочетает декларативное описание знаний с элементами процедурного [2].

Действительно, прохождение по ветвям графа моделирует процесс преобразования данных в информацию. Граф преобразования представляется в виде следующего множества

$$G_1^D = \{V_1^D, E_1^D\}, \quad (5)$$

где $V_1^D = \{v_j\}$ – множество вершин (nodes) графа,

D – данные,

I – информация,

v_j – j -я вершина графа преобразования данных в

информацию G_1^D ,

$E_1^D = \{e_k\}$ – множество дуг (edges) графа G_1^D .

Каждая дуга графа e_k , соединяющая вершины графа v_i и v_j , при построении его алгебрологической модели описывается бинарным предикатом вида:

$$P_k(x_i, x_j) = 1, \quad (6)$$

где x_i, x_j атрибуты вершин графа v_i и v_j , которые представляют собой символьные переменные, заданные на областях определения данных в вершинах графа.

Дуги графа которые заданы предикатами алгебры конечных предикатов, обладают следующей особенностью. Предикаты из множества $\{P_k\}$, $k = \overline{1, m}$, где m – число дуг графа, задаются на множестве допустимых значений, которое представляется произвольным набором символов. Поэтому такой предикат может задать условное действие по преобразованию, которое моделируется дугой графа

Механизм условных дуг графа можно представить следующим фрагментом, включающим явные I_1 и I_2 , а также неявную I_1 связи:

Приведенная выше условная связь I_1 , также описываемая бинарным предикатом, при ее задействовании исключает влияние данных из вершины p_2 и исключает информацию I_2 из результирующего знания.

Такие условные связи $\{I'_j, j = \overline{1, l}\}$ могут быть получены на основании вычисления систем (4) (вся формализованная информация о данных или явные знания) и (5) (вся информация о данных – формализованная и неформализованная, или сумма явных и неявных знаний).

Результатом данного преобразования является статическая информационная структура.

Второе преобразование – информации в явные знания, в качестве исходных данных использует полученную ранее информационную структуру I , а также дополнительные знания. Отметим эту деталь: преобразования информации в знания требует использования дополнительных явных знаний и, в случае естественного интеллекта, сопровождается влиянием неявных знаний.

Процесс преобразования должен включать в себя предварительную фильтрацию информации, как показали исследования в области естественного интеллекта [3].

Такая фильтрация выполняется на основании шаблона, задаваемого явными знаниями [4]. Шаблон фильтрации (по приведенным выше причинам) в общем виде целесообразно представить в виде конъюнкции предикатов:

$$\bigwedge_{i=1}^k I_i \wedge \bigwedge_{j=1}^l I'_j, \quad i = \overline{1, k}, \quad j = \overline{1, l}. \quad (7)$$

Входами фильтрующей реляционной сети являются узлы p_1 и p_2 . Задание значений переменных x_1 и x_2 этих узлов на нулевом такте работы реляционной сети определяет дальнейшую цепочку вычислений данных в узлах сети. Выходы фильтрующей сети в примере – это узлы p_3 , p_6 и p_7 .

Пусть при вычислении выражения (7) при заданных значениях переменных $p_1 - p_8$ предикаты Γ'_1 и Γ'_3 , описывающие две из трех неявных связей, равны 0. Это говорит о том, что информация, формально представленная предикатами Γ'_1 и Γ'_3 , противоречит совокупности информации, представленной предикатами $I_1 - I_7$ и Γ'_2 . Эта информация соответствует набору данных, заданному значениями переменных $p_1 - p_8$. В этом и состоит работа фильтрации информации.

Отфильтрованная информация преобразуется в знания на основе паттернов (явных знаний) [5] и неявной составляющей естественного интеллекта, что можно представить моделью в виде предиката второго порядка:

$$M^*(I_1, I_2, \dots, I_8, \Gamma'_2) \quad (8)$$

Наиболее важный с практической точки зрения случай предиката M^* представляет собой решение системы предикатных уравнений для заданных начальных условий.

Полученную систему можно представить в виде графа реляционной сети. В модели Γ'_1 и Γ'_3 принимают нулевые значения, поэтому исключаем их из графа.

Данное преобразование выполняется при решении многих задач искусственного интеллекта, в частности при интеллектуальном анализе данных, процессов, веб, социальных сетей.

Результатом данного преобразования являются явные знания, которые, фактически, представляют собой некоторые паттерны на информации и, в отличие от информации, имеют динамическую природу:

$$M^*(I_1, I_2, \dots, I_7, \Gamma'_2), \quad (9)$$

Рассмотрим этапы получения явных знаний.

1). Вначале мы имеем только неявные знания, их носителями являются люди. Обозначим неявные знания $\text{Imp}(p_1, p_2, \dots, p_n)$.

2). Неявные знания не формализованы, но мы можем получить информацию $I_i(p_1, p_2, \dots, p_n)$, $I_i, i = \overline{1, k}$, проведя структуризацию неявных знаний. При этом используются формализованные и не формализованные коллективные знания.

3). Если заданы конкретные условия задачи, можно на основании полученной информации вычислить соответствующие этим условиям данные.

4). Эти данные могут быть получены другим

человеком и проинтерпретированы с использованием коллективных знаний. В результате получим новую информацию.

5). Применяя различные шаблоны, человек может формировать неявные и явные знания.

Сравнив фрагменты явных знаний K_E^* и K_E^{**} об одном предмете, явлении, процессе и выделив отличия, мы получаем возможность формализовать неявную составляющую знаний, что и составляет сущность третьего преобразования предлагаемой модели:

$$\forall (K_E^*, K_E^{**}) \exists K_T, P_{KT}^{KE} = K_E^* \text{ XOR } K_E^{**}. \quad (10)$$

Результатом данного преобразования являются формализованные неявные знания, которые представляются в виде набора предикатов, формализующих разность между исходными фрагментами явных знаний:

$$K_T = \begin{cases} I_{1'}, \\ I_{2'}, \\ \dots \\ I_{n'}, \\ \Gamma'_{1'}, \\ \Gamma'_{2'}, \\ \dots \\ \Gamma'_{m'}, \end{cases} \quad (11)$$

где $I_{i'}$, $\Gamma'_{j'}$ – бинарные предикаты, формализующие соответственно явную и неявную информацию. Индексы при предикатах имеют штрихи, чтобы указать, что нумерация предикатов $I_{i'}$, $\Gamma'_{j'}$ во фрагментах знаний K_E^* , K_E^{**} и не совпадает. Кроме того, фрагменты знаний K_E^* , K_E^{**} семантически связаны (описывают связный процесс или объект), а набор предикатов – нет. В набор K_T попадают предикаты из K_E^* , отсутствующие в K_E^{**} и из K_E^{**} , отсутствующие в K_E^* .

Обобщив изложенные преобразования, получаем обобщенную предикатную модель преобразования данных в информацию, информации в явные знания, и, в итоге, явных знаний в формализованные неявные.

- 1) $I(p_1, p_2, \dots, p_n)$;
- 2) $I(p_1, p_2, \dots, p_n) \rightarrow \{I_i(p_j, p_k), i = \overline{1, m}\}$;
- 3) $M(I_1, I_2, \dots, I_m)$;
- 4) $M^*(I_1, I_2, \dots, I_m, \Gamma'_{1'}, \Gamma'_{2'}, \dots, \Gamma'_{l'})$;
 $\{I_i, i = \overline{1, m}; \Gamma'_{j'}, j = \overline{1, l}\}$;
- 5) $K_E^*, K_E^{**} \subseteq M^*, P_{KT}^{KE} = K_E^* \text{ XOR } K_E^{**}$;
- 6) $K_T = F(K_E^*, K_E^{**})$.

Первая часть модели – формирование информации по данным.

Вторая – создание реляционной сети (бинаризация предиката I).

Третья – формализация явных знаний.

Четвертая – формализация явных и неявных знаний.

Пятая – построение двоичной матрицы разности между эталонным K_E^* и реальным K_E^{**} фрагментами явных знаний.

Шестая – формализация неявных знаний.

Рассмотрев все трансформации данных в знания, мы можем уточнить характеристики наборов исходных данных, которые могут быть использованы в системах искусственного интеллекта для выявления неявных знаний.

Исходными данными для модели является набор событий D , отражающий особенности и поведение анализируемых объектов, процессов в предметной области.

Каждое событие из набора обладает рядом атрибутов, в частности метками времени t , значимости v , связанного с ним объекта o , а также человека-исполнителя e , который и сформировал данное событие:

$$D = (t, v, o, e). \quad (13)$$

Результатом использования предложенной обобщенной модели является представление неявных знаний в виде (13).

Выводы

Предложена обобщенная предикатная модель процесса выявления неявных знаний, которая основана на процессе преобразования данных в информацию, а информации в знания и выделяет стадию

использования явных знаний для выделения неявных зависимостей.

Предложенная модель, как уже отмечалось выше, позволяет разделить явные и формализованные неявные знания по форме их представления. Явные знания, которые получены в результате второго преобразования модели, имеют декларативную форму и представляют собой набор паттернов (шаблонов), которые описывают структуру, поведение объектов, а также поведение процессов изучаемой предметной области.

Список литературы

1. Ackoff, Russell. *From Data to Wisdom / Journal of Applied Systems Analysis* 16: P. 3–9.
2. О мозгоподобных ЭВМ / М.Ф.Бондаренко, З.В. Дударь, В.А. Лецинский, И.А. Лецинская, С.Ю. Шабанов-Кушнаренко // *Радиоэлектроника и информатика*. – 2004. – № 2. – С. 89-105.
3. Румизен Мелисси Клеммонс. *Управление знаниями ... / Мелисси Клеммонс Румизен*. – М.: АСТ: Астрель, 2004. – 318 с.
4. Нонака И. *Компания-создатель знания. Зарождение и развитие инноваций в японских фирмах [Текст] / И. Нонака, Х. Такеучи*. – М., 2003. – 384 с.
5. Tsuchiya S. "Improving knowledge creation ability through organizational learning," in *ISMICK 1993: Proceedings of the International Symposium on the Management of Industrial and Corporate Knowledge*, 1993. – Pp. 87-95.

Поступила в редколлегию 27.06.2014

Рецензент: д-р техн. наук, проф. С.Ф. Чалий, Харьковский национальный университет радиоэлектроники, Харьков.

УЗАГАЛЬНЕНА ПРЕДИКАТИВНА МОДЕЛЬ ПРОЦЕСУ ВИЯВЛЕННЯ НЕЯВНИХ ЗНАТЬ В ІЄРАРХІЇ «ДАНІ-ІНФОРМАЦІЯ-ЗНАННЯ»

С.Ю. Шабанов-Кушнаренко, Ю.С. Новіков

Вдосконалений багаторівневий підхід до виділення неявних залежностей «дані – інформація – явне знання – неявне знання». Пропонований підхід формально представлений у вигляді предиката, що відображає процесний характер формування і трансформації знання. Запропонована узагальнена предикативна модель процесу виявлення неявних знань, яка заснована на процесі перетворення даних в інформацію, а інформації в знання і виділяє стадію використання явних знань для виділення неявних залежностей.

Ключові слова: алгебра скінченних предикатів, реляційна мережа, знання.

THE OF GENERALIZED PREDICATIVE MODEL OF PROCESS OF EXPOSURE OF NON - OBVIOUS KNOWLEDGE IS IN THE HIERARCHY «DATA-INFORMATION-KNOWLEDGE»

S.Yu. Shabanov-Kushnarenko, Yu.S. Novikov

The multilevel going is improved near the selection of non-obvious dependences «data – information – obvious knowledge – non-obvious knowledge». Offered approach formally presented as a predicate reflecting process character of forming and transformation of knowledge. The generalized predicate model of process of exposure of non-obvious knowledge is offered, that is based on the process of conversion of data in information, and to information in knowledge and distinguishes the stage of the use of obvious knowledge for the selection of non-obvious dependences.

Keywords: algebras of finite predicates, relational network, knowledge.