И. Е. Мазурок

ФОРМАЛЬНЫЕ ПРАГМАТИЧЕСКИЕ ОНТОЛОГИИ В МОДЕЛЯХ СОДЕРЖАНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Современные распределенные информационные системы характеризуются огромными объемами хранимой информации и широким (свободным) доступом к ней. Для анализа информационных свойств таких систем разрабатываются различные по эффективности и используемому аппарату математические модели [2]. Важной особенностью этих систем является их априорная интеллектуализация и наличие проблем информационной безопасности [4].

Сложившаяся ситуация порождает необходимость оптимизации системы использования данных ресурсов с учетом прагматического профиля пользователя. Создание предпосылок подобной оптимизации за счет разработки специального вида формальных когнитивно-структурных моделей является целью данной работы. Для достижения цели исследования необходимо решение задач формирования системы формальных понятий и разработки прагматической онтологии на основе прагматических профилей пользователей.

Для моделирования содержания информационных ресурсов в рамках их использования в составе интеллектуальной информационной системы необходимым является построение системы понятий в виде полной онтологии. В настоящее время разработаны системы различных онтологий [1, 6]. В их основе лежат системы таксономий, разработанные людьми — конкретными специалистами. В то же время развивается направление фолксономий [5], являющихся результатом труда практически неограниченного числа обычно анонимных пользователей по созданию т. н. «облаков тэгов». К сожалению, создание всеобщей полной таксономии в рамках какой-либо онтологии или даже системы онтологий не представляется возможным в виду субъективности,

противоречивости и недостоверности имеющихся информационных ресурсов.

Таким образом, одной из наших задач будет создание систем ***псевдопонятий*** на основе сложившихся субъективных комплексов представлений. В отличие от формальных понятий, фигурирующих в технике *Formal Concept Analysis* (FCA), ***формальные псевдопонятия*** характеризуются только своим объемом и контекстом:

$$^{p}C = \langle V, Discourse \rangle . \qquad \textbf{(1)}$$

Обобщенным дискурсом будем называть информационную категорию, представляющую собой зафиксированное в материальной форме структурное объединение коммуникативных единиц из произвольного набора языков или знаковых систем, допускающее осмысленное толкование в определенном контексте. Для такого дискурса характерно объединение текстового, аудиовизуального контекста с конструкциями общих и специальных знаковых систем и языков (математические формулы, графовые модели, эмоциональные сигнатуры, фрагменты музыкальных партитур и пр.). Характерной чертой обобщенного дискурса является его смысловая и структурная целостность, позволяющая рассматривать его как цельное информационное образование. Примерами обобщенного дискурса могут служить современные научно-технические статьи, отдельные разделы мультимедийных энциклопедий, схемы ***MindMap*** и др.

Введем также понятие ***дискурсной базы*** онтологии — множества обобщенных дискурсов

$$\Theta = \left\{ \theta_i \,\middle|\, i = \overline{1, n_\theta} \right\}. \qquad \textbf{(2)}$$

Выделим ***базовое множество*** псевдоонтологии — множество базовых объектов (***концептов***) для ее построения (слова, муземы, сэмплы, изображения или их фрагменты, знаки и т.п., фигурирующие в соответствующем обобщенном дискурсе)

$$\Lambda = \left\{ \lambda_i \,\middle|\, i = \overline{1, n_\lambda} \right\}. \qquad \textbf{(3)}$$

Базовое множество Λ в нашей ситуации является индуцированным понятием, которое строится на основе конкретной дискурсной базы Θ, поскольку в задачах моделирования и анализа наполнение системы является первичным. То есть элементы базового множества формируются в процессе структурного анализа дискурсной базы (множество текстов, Интернет, содержательная часть интеллектуальной системы и т. п.). Это позволяет сформулировать следующие задачи.

Задача 1. Построение базового множества Λ (2) на основе структурного анализа дискурсной базы Θ (3).

Отсутствие семантических объектных отсылок позволяет решать задачу алгоритмическими методами синтаксического и структурного анализа элементов дискурсной базы.

Задачи построения онтологий связаны с определением категории. Будем выделять ***категории (формальные псевдопонятия) 1-го порядка*** ${}^1\mathrm{K}$, оперирующие непосредственно элементами базового множества.

$$
\begin{aligned}
&{}^1k = \left\langle {}^1\mathrm{B}(\Lambda), \Sigma \right\rangle, \\
&{}^1\mathrm{B}(\Lambda) \in 2^{\Lambda}, \\
&\Sigma : \begin{cases} 2^{\Lambda} \to \Lambda \cup \Theta, \\ 2^{\Lambda} \to N, \end{cases}
\end{aligned}
\qquad \textbf{(4)}
$$

где ${}^1\mathrm{B}(\Lambda)$ задает объем формального понятия, а Σ представляет собой именующую или идентификационную функцию. Таким образом, ***формальное псевдопонятие*** есть отраженный в содержательной части интеллектуальной системы образ реального понятия субъекта (субъектов) именованный субъектом или алгоритмически идентифицируемый самой системой. Данный термин приходится вводить для того, чтобы не использовать семантические референции за пределы формальной системы, что сделало бы невозможным автоматизированный анализ.

Существенно, что область значений (обозначения понятий) функции Σ в именующей формулировке лежит в области концептов и дискурсов. Последнее является его неалгоритмическим определением. В более общем случае имеется некоторое

множество таких концептов или дискурсов, что создает «облако определений», аналогичное облаку тэгов для web-ресурсов [7]. Это позволяет выполнить структурную детализацию именования категорий, выделяя имя понятия (name), его определение (def) и уникальный идентификатор (id):

$$\Sigma = \Sigma\left\langle def, name, id\right\rangle \begin{cases} def : 2^{\Lambda} \to \Lambda \cup \Theta, \\ name : 2^{\Lambda} \to \Lambda, \\ id : 2^{\Lambda} \to N. \end{cases} \qquad \textbf{(5)}$$

На основании изложенного можно сформулировать задачу построения ***формальной псевдокатегоризации 1-го порядка***:

$$^{1}\mathrm{K} = \left\{ {}^{1}k = \left\langle {}^{1}\mathrm{B}\left(\Lambda\right), \Sigma = \left\langle def, name, id\right\rangle \right\rangle \right\}. \qquad \textbf{(6)}$$

Задача 2. Построение системы формальных понятий первого порядка $^{1}\mathrm{K}$ на основании алгоритмического структурного анализа, анализа эмпирических фолксономий [5] и обычных онтологий.

На первое место в этом анализе следует поставить методы алгоритмического структурного анализа, поскольку стихийно сложившиеся массивы данных значительно превышают ограниченный объем фолксономий и сравнительно небольшой объем профессиональных онтологий.

Ключевым методом анализа на данном этапе является контекстный семиотический анализ шаблонов употребления элементов базового множества Λ на основе идей FCA [138]. Разовьем эти идеи для нашего случая.

Исходя из априорных данных о характере элементов обобщенного дискурса, выделяются ***ролевые синтаксические схемы***. Каждая синтаксическая схемы задается предикатом. Ролевые синтаксические схемы ***1-го порядка***

$$^{1}Schema = \left\{ {}^{1}P_{i}^{(n_i)}\left(x_1, x_2, \cdots, x_{n_i}\right) \right\} \qquad \textbf{(7)}$$

задают синтаксические отношения для элементов обобщенного дискурса.

Ролевые синтаксические схемы ***2-го порядка*** в качестве аргументов получают множества

$$^{2}Schema = \left\{ {}^{2}P_i^{(n_i)}\left(x_1, x_2, \cdots, x_{n_i}\right) \middle| x \in 2^{\Lambda} \right\} \qquad \textbf{(8)}$$

и задают более общие синтаксические отношения между множествами объектов.

Если схемы задаются средствами нечеткой логики, то соответствующие предикаты принимают вид

$$\begin{aligned} &{}^{m}P_i^{(n_i)}\left(x_1, x_2, \cdots, x_{n_i}\right): \Lambda^{n_i} \to [0;1], \\ &{}^{m}P_i^{(n_i)} \in {}^{m}Schema, \\ &m \in \{1,2\}. \end{aligned} \qquad \textbf{(9)}$$

Синтаксические схемы вводятся таким образом, чтобы отразить правила построения высказываний языка, на котором формулируются утверждения обобщенного дискурса. Поскольку в состав большинства интеллектуальных информационных систем входят разнородные документы, задача разработки синтаксических схем тесно связана с формализацией синтаксиса языков представления.

Далее формируется таблица формального контекста [6]. От общепринятой FC -таблицы она отличается тем, что в ней и сроки и столбцы помечены объектами из Λ, а не концептами и их атрибутами Построение таблицы в случае использования схем (7) выглядит так:

$$^{1}FC^{i}_{j_1, j_2, \cdots, j_{n_i}} = {}^{1}P_i^{(n_i)}\left(x_{j_1}, x_{j_2}, \cdots, x_{j_{n_i}}\right). \qquad \textbf{(10)}$$

Наиболее типичный пример — предикаты вида $^{1}P\left(x, y\right)$, для которых выполняется условие $\forall x, y \in \Lambda : {}^{1}P\left(x, y\right) = {}^{1}P\left(y, x\right)$. В этом случае мы получаем обычную двумерную симметричную таблицу.

Если используются синтаксические схемы *2*-го порядка (8), то производится предварительная агрегация концептов, доставляющих истину соответствующему предикату. Наиболее употребительным оказался случай одноместных предикатов. Для него

$$ {}^{2}FC^{i}_{j_1,j_2} = \begin{cases} 1, \forall j_1, j_2 : x_{j_1} \in X, x_{j_2} \in X, {}^{2}P^{(2)}_{i}(X) = 1; \\ 0, \text{if else.} \end{cases} \qquad \textbf{(11)} $$

Для учета частотной значимости контекста или при использовании элементов нечеткой логики (9) для задания предикатов таблицы следует представить в виде

$$ {}^{2}FC^{i}_{j_1,j_2} = \sum_{\substack{\forall j_1, j_2: \\ x_{j_1} \in X, \\ x_{j_2} \in X}} {}^{2}P^{(2)}_{i}(X). \qquad \textbf{(12)} $$

Дальнейшая категоризация проходит по модифицированному *FCA*-алгоритму. Основная модификация касается введения коэффициента включения ψ, который задает минимальное относительное значение мощности категорий, при котором еще производится выделение понятий (т.н. процедура формирования исключений). Легко показать, что получаемые по этому алгоритму структуры также являются решетками.

Категоризация, построенная в результате решения задачи 2, является ***псевдокатегоризацией,*** поскольку никакие дополнительные семантические условия на понятия в категории не накладываются. Объемы различных понятий могут перекрываться, входить один в другой и даже полностью совпадать. Такую свободу приходится предоставлять ввиду того, что стихийно складывающиеся в содержательной части интеллектуальной системы формальные понятия (в отличие от специально создаваемых средствами языков описания метаданных) не имеют каких-либо логических ограничений.

Построение моделей формальных категорий более высоких порядков вызывает определенные затруднения, связанные со свободой стихийного формирования классов. Начнем с определения категории 2-го порядка

$$ \begin{aligned} &{}^{2}k = \left\langle {}^{2}\mathrm{B}(\Lambda), \Sigma \right\rangle, \\ &{}^{2}\mathrm{B}(\Lambda) \in 2^{\Lambda \cup {}^{1}\mathrm{K}}. \end{aligned} \qquad \textbf{(13)} $$

На первый взгляд, объем понятия вводится несколько вольно — класс ${}^{2}\mathrm{B}(\Lambda)$ содержит как множества, так и отдельные объекты. На практике это не создает проблем, благодаря введению конкретных единичных понятий — множеств из одного элемента для всех элементов базового множества. Для конкретных единичных понятий мы будем использовать термин ***категорий 0-го порядка***

$$\begin{aligned} {}^{0}k &= \left\langle {}^{0}\mathrm{B}(\Lambda), \Sigma \right\rangle, \\ {}^{0}\mathrm{B}(\Lambda) &= \{\lambda_i\}, \\ \Sigma\left({}^{0}k\right) &= \lambda_i. \end{aligned} \qquad \textbf{(14)}$$

По аналогии можем определить ***формальные понятия n-го порядка***

$$\begin{aligned} {}^{n}k &= \left\langle {}^{n}\mathrm{B}(\Lambda), \Sigma \right\rangle, \\ {}^{n}\mathrm{B}(\Lambda) &\in 2^{\bigcup_{j=0}^{n-1}\left({}^{j}\mathrm{B}(\Lambda)\right)} \end{aligned} \qquad \textbf{(15)}$$

и (в предельном переходе) просто — ***формальные понятия***

$$\begin{aligned} k &= \left\langle \mathrm{B}(\Lambda), \Sigma \right\rangle, \\ \mathrm{B}(\Lambda) &\in 2^{\bigcup_{j=0}^{\infty}\left({}^{j}\mathrm{B}(\Lambda)\right)}. \end{aligned} \qquad \textbf{(16)}$$

К счастью, на практике нам не придется работать с такого рода булеанами, поскольку $\mathrm{B}(\Lambda)$ всегда будут оставаться конечными структурами. В реальной конечной дискретной реализации хранения содержательной части интеллектуальной информационной системы не может возникнуть бесконечно высокая иерархия. В то же время потенциальная возможность возникновения циклов в таких стихийных определениях не позволяет отказаться от (16), по крайней мере, на стадии анализа.

Наконец, общая задача построения ***формальной псевдокатегоризации***

$$O = \left\{ k = \left\langle \mathrm{B}(\Lambda), \Sigma = \left\langle def, name, id \right\rangle \right\rangle \right\} \qquad \textbf{(17)}$$

должна привести к созданию формальной псевдоонтологии.

Задача 3. Построение полной системы формальных понятий K на основе алгоритмического структурного анализа и анализа эмпирических фолксономий [5] и профессиональных онтологий.

Для моделирования «контента» алгоритмическими методами может быть создана формальная псевдоонтология O, являющаяся отражением знаний, эмпирически накопленных в составе содержательной части интеллектуальной системы.

Ближайшая цель дальнейших исследований лежит в области изучения собственных свойств когнитивных структур интеллектуальных информационных систем [8] на основе семиотических моделей и разработанной системы формальной онтологии. Некоторые результаты, полученные в этой области [1, 9], позволяют надеяться на перспективность работы.

Задача 4. Формирование модели целевого ***информационного поведения*** пользователя информационной системы.

Для решения задачи необходимо проведение долговременного наблюдения и журнализации действий пользователя в процессе взаимодействия с информационной системой. Анализируется перечень затребованных информационных ресурсов с указанием даты и времени. Для каждого зафиксированного акта информационного поведения $\beta_i(h)$ фиксируется ресурс $\theta \in \Theta$ (2) и задается последовательность штампов даты-времени для начала t_{start} и конца t_{end} каждого фрагмента работы пользователя с ним. Полная база $\mathrm{B}(h)$ актов информационного поведения пользователя h совместно с дескрипторами ресурсов образует нормативный дискурс пользователя. ***Анонимная общая база информационного поведения*** B формируется объединением обезличенных актов информационного поведения:

$$\begin{aligned} &\mathrm{B} = \bigcap_h \mathrm{B}(h), \\ &\mathrm{B}(h) = \left\{ \beta_i(h) \middle| i = \overline{1, n_\beta(h)} \right\}, \\ &\beta_i(h) = \langle \theta, t_{start}, t_{end} \rangle. \end{aligned} \qquad \textbf{(18)}$$

Анонимность общей базы информационного поведения означает отсутствие личной информации реального субъекта

информационного поведения. Иными словами, субъект в рамках общего прагматического анализа формальной онтологии фигурирует как виртуальный образ, не имеющий уникального персонального идентификатора.

Задача 5. Формирование прагматических профилей пользователя и системы дескрипторов, позволяющих опознавать текущий профиль.

Во время работы с информационной системой пользователь $h \in \mathrm{H}$ в различное время решает различные комплексы поисковых задач, определяемых целью. Цель или связанную группу персональных целей будем определять через ***персональный прагматический профиль*** $\pi_i(h)$, а полную совокупность прагматических профилей пользователя — как ***персональную прагматику*** $\pi(h) = \left\{ \pi_i(h) \middle| i = \overline{1, n_\pi(h)} \right\}$.

Для построения прагматических профилей пользователя имеются два подхода, позволяющих построить различные с точки зрения дальнейшего применения виды прагматик. Это, во-первых, прагматический профиль $\pi_i(h)$ пользователя h:

$$\Pi(h) = \left\{ \pi_i(h) \middle| i = \overline{1, n_\pi(h)} \right\}. \qquad \textbf{(19)}$$

Во-вторых, используя полученное кластеризацией множество не персональных ***типовых прагматических профилей***

$$\aleph(\mathrm{B}) \to \left(\Pi = \left\{ \pi_i \middle| i = \overline{1, n_\pi} \right\} \right), \qquad \textbf{(20)}$$

можно соотнести каждый акт информационного поведения пользователя системы с определенным типовым профилем. Результирующая прагматическая модель в этом случае представляет собой нечеткое множество

$$\tilde{\Pi}(h) = \left\{ \langle \pi_i, p_i \rangle \middle| i = \overline{1, n_\pi} \right\}, \quad p \in (0;1), \qquad \textbf{(21)}$$

где p_i — характеристика степени вхождения типовой прагматики π_i в прагматику $\tilde{\Pi}(h)$ пользователя. Прагматику $\tilde{\Pi}(h)$ будем называть ***социальной прагматикой*** пользователя.

Задача 6. Формирование прагматической формальной онтологии как совокупности формальных понятий, определяемых базой онтологии и текущим прагматическим профилем.

Предложенный подход к решению задач моделирования содержания сверхбольших информационных репозиториев на базе системы формальных прагматических онтологий позволяет создавать эффективные средства анализа и поиска информации. Одним из очевидных применений предложенного подхода является создание адаптивных интерактивных систем для работы в Web-средах.

---

1. Артемьева И. Л. Многоуровневые математические модели предметных областей // Искусственный интеллект. — Т. 4, 2006. — С. 85—94.
2. Мазурок І. Є. Формальний опис класу інформаційних систем, заснованих на знаннях // Збірник статей: Нові інформаційні технології навчання в учбових закладах України, 1998. — Випуск 6. — С. 217—226.
3. Мазурок И. Е. Метамоделирование содержательной части систем, основанных на знаниях// Моделирование в прикладных научных исследованиях.– Одесса: Од. гос.политехн. ун.-т. — 1998. — С. 51—55.
4. Мазурок І. Є. Інтелектуальні інформаційні системи з розвиненою моделлю захисту і авторизації // Вісник Запорізького державного університету. Сер. фіз.-мат. наук. — 2002. — № 1. — С. 65—68.
5. Мазурок І. Є. Модель реалізації збереження когнітивних структур інтелектуальних інформаційних систем // Вісник Львівського університету. Серія: Прикладна математика та інформатика. — 2004. — Вип. 9. — С. 96—105.
6. Andreas Hotho, Robert Jдschke, Christoph Schmitz, Gerd Stumme Information. Retrieval in Folksonomies: Search and Ranking// The Semantic Web: Research and Applications. — Volume 4011/2006, pp. 411—426.
7. Cristani M., Cuel R. A Survey on Ontology Creation Methodologies.// Semantic Web & Information Systems. — 2005. — 1(2). Pp.48—68.
8. Haynes L., Selcukoglu A., Sunah Suh, Karahalios K. Tagscape: Navigating the Tag Landscape// Human-Computer Interaction. – Volume 4663/2007, pp. 264—267.
9. Uta Priss. Formal Concept Analysis in Information Science //Annual Review of Information Science and Technology. Vol 40, 2006, pp. 521—543.