УДК 004.65

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПРОЦЕССА ИНТЕГРАЦИИ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ ОНТОЛОГИЙ

Бубарева О. А., Попов Ф. А.

Бийский технологический институт (филиал) федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Алтайский государственный технический университет им. И. И. Ползунова» (Россия,659305, Алтайский край, г. Бийск, ул. Трофимова, 27) angel@bti.secna.ru

Для обеспечения автоматизации и информатизации всех видов деятельности ВУЗы разрабатывают интегрированные автоматизированные информационные системы (ИАИС). При постоянных изменениях в бизнес-процессах вуза разработчики ИАИС вынуждены постоянно заниматься корректировкой программ и моделей данных, что приводит к структурной и семантической неоднородности интегрируемых систем.

С целью разрешения данной проблемы в статье предлагается математическая модель процесса интеграции ИС с неоднородными онтологическими спецификациями, позволяющая анализировать семантические связи, закономерности и зависимости, возникающие между ними. Предложен метод определения меры семантической близости концептов (классов объектов) как суммы атрибутивной, таксономической и реляционной составляющих с учетом весовых коэффициентов. С целью автоматического определения весовых коэффициентов используется генетический алгоритм. Предложен также метод классификации уровней близости концептов с целью построения результирующей (интегрированной) онтологии из нескольких исходных. Представлены результаты вычислительного эксперимента, подтверждающие целесообразность построения такого рода моделей и их практическую значимость.

Ключевые слова: онтология, интеграция данных, информационная система, семантическая близость.

MATHEMATICAL MODEL OF THE INTEGRATION OF INFORMATION SYSTEMS BASED ONTOLOGY

Bubareva O. A. Popov F. A.

Biysk Technological Institute (branch) of the federal government budget of educational institutions of higher education "Altai State Technical University. of I. I. Polzunov" (Russia,659305, Altay territory, Biisk, street Trofimova, 27) angel@bti.secna.ru

Universities develop the integrated automated information systems (IAIS) for supporting automation and information of all types of activity. With continuous changes in business processes of university, developers have to constantly deal with IAIS adjustment programs and data models, which leads to structural and semantic heterogeneity of integrated systems.

In order to solve this problem in the paper we propose a mathematical model of the integration of IS with heterogeneous ontological specifications, which analyzes the semantic context, patterns and dependencies that arise between them. A method for determining the measure of semantic similarity of concepts (object classes) as the sum of the attribute, taxonomic and relational component of the weights. In order to automatically determine the weights used by a genetic algorithm. Proposed as a method of classification levels of similarity of concepts in order to build the resulting (integrated) ontology from multiple source. The results of computing experiment confirming expediency of creation of such models and their practical importance are presented.

Keywords: ontology, integration of data, information system, semantic similarity.

Для обеспечения автоматизации и информатизации всех видов деятельности вузы разрабатывают интегрированные автоматизированные информационные системы (ИАИС) [2]. При постоянных изменениях в бизнес-процессах вуза разработчики ИАИС вы-

нуждены постоянно заниматься корректировкой программ и моделей данных, что приводит к структурной и семантической неоднородности интегрируемых информационных систем и необходимости повторной разработки приложений-конверторов. Для решения проблемы семантической неоднородности информации возможно использование онтологий [3]. Создание общедоступных онтологий предметных областей в определённой мере решает проблему неоднородности онтологических спецификаций для определённых групп ИС. Однако в ИАИС вуза используются несколько идентифицированных предметных областей, к которым предъявляются различные требования. Поэтому, для обеспечения семантически корректной интероперабельности неоднородных ИС, в контексте предметной области задачи, необходимо выяснить общность и различия онтологий, лежащих в их основе, согласовать неоднородные онтологические спецификации и на базе соответствий онтологических контекстов осуществлять преобразование информации [1].

Авторами разработан алгоритм построения результирующей онтологии из нескольких исходных на основе результатов сравнения концептов, отношений и атрибутов. Задача интеграции ИС сводится к задаче построения отображений и интеграции онтологий, а затем и установление взаимосвязей схем интегрируемых ИС, т.е. сохранение соответствия множества онтологий ИС заданному набору семантических зависимостей, позволяя установить взаимодействие между ИС.

Как правило, объектная схема ИАИС вуза включает в себя элементы, которые соответствуют сущностям разных предметных областей, каждый объект характеризуется значениями набора атрибутов и представляется как множество упорядоченных пар вида

$$u = \{< a_i, d_i >\}, \quad (1)$$

где a_i – атрибут объекта, d_i – значение атрибута $i \in [1 \dots n]$, n – количество атрибутов.

Базовым понятием предлагаемой модели является концепт C. Каждый концепт онтологии информационной системы идентифицируется по имени и характеризуется типом. Поэтому концепт зададим как:

$$C_i = (Name_i, type_i), \quad (2)$$

где $Name_i$ – уникальное имя (идентификатор) -го концепта; $type_i$ – тип i-го концепта (абстрактный, представимый, либо составной).

Зададим следующее множество концептов $C = \{C_i | i = 1,2, \dots, n\}$ и множество отношений между концептами:

$$R = \{R_1, R_2, R_3\}, \quad (3)$$

где R_1– отношение наследования (отношения «класс-подкласс»), $R_1(C_1, C_2)$, где C_1 – надкласс концепта C_2;

R_2 – отношение агрегации (отношения «часть/ целое»), $R_2(C_1, A')$: атрибуты концепта C_1 входят во множество атрибутов всех концептов A'.

R_3 – отношение ассоциации (семантические отношения), обладающее свойством транзитивности.

Вводится функция интерпретации I, сопоставляющей каждому концепту онтологии множество элементов объектной схемы информационной системы, и каждой роли – декартово произведение таких множеств. Интерпретация называется моделью онтологии $O(I \epsilon M(O))$, если она удовлетворяет всем значениям в C и R. Онтология, не имеющая моделей, называется противоречивой.

Описание онтологических моделей информационных систем, автоматизирующих деятельность ВУЗа, которые состоят из информационных объектов, формально можно представить в следующем виде:

$$O =< C, A, G, M_A, M_c, R, I >, \quad (4)$$

где $C = \{C_i | i = 1,2, \dots, n\}$ – множество концептов; $A = \{a_{ij} | ij = 1,2, \dots, j\}$ – множество атрибутов концептов; $G = \{g_{ik} | ik = 1,2, \dots, k\}$ –множество ограничений, накладываемых на атрибуты; $M_c: C \to 2^A$ – отображение, задающее для каждого концепта множество его атрибутов; $M_A: A \to G$ – отображение, задающее ограничения на каждый атрибут; *R* – множество отношений; *I* – функция интерпретации.

Информационная система, использующая онтологию *O*, представлена в виде:

$$U^O =< O, U, M_U, M_R >, \quad (5)$$

где $U = \{u_1, u_2, \dots, u_n\}$ – множество элементов объектной схемы ИС; $M_U: U \to C$ – отображение, ставящее в соответствие элементу объектной схемы его концепт, $M_R: U \times U \to R$ – отображение, ставящее в соответствие связям между элементами объектной схемы их отношения в онтологии, и для любого элемента $u \in U$ выполняется условие: множество атрибутов элемента объектной схемы u соответствует атрибутам его концепта, т.е. $\{a: < a, d > \in u\} = M_c(M_U(u))$.

Обозначим через H^O– множество онтологических моделей информационных систем, использующих онтологию O.

Обозначим изменение информационной системы как отображение:

$$F: H^O \to H^O, \tag{6}$$

где H^O – множество неоднородных информационных систем.

Изменение онтологии:

$$U^{\bar{O}} = \{U_1^{O^1}, U_2^{O^2}, \dots, U_N^{O^N}\}, \tag{7}$$

где $U_i^{O^i} =< O_i, U_i, M_{U_i}, M_{R_i} >$ и $O_i =< C_i, A_i, G_i, M_{C_i}, M_{A_i}, R_i, I_i >$, и введем обозначения: $\bar{C} = \bigcup_{1 \le i \le N} C_i$, $\overline{R} = \bigcup_{1 \le i \le N} R_i$, $\overline{I} = \bigcup_{1 \le i \le N} I_i$, $\bar{A} = \bigcup_{1 \le i \le N} A_i$, $\bar{G} = \bigcup_{1 \le i \le N} G_i$, $\overline{U} = \bigcup_{1 \le i \le N} U_i$.

Различные онтологии ИС, входящие в O, могут иметь пересекающиеся множества атрибутов, типов и концептов. На базе нескольких исходных онтологий, которые используют информационные системы, осуществляется построение результирующей онтологии с сохранением исходных спецификаций в таком виде, чтобы она включала все возможные отношения между концептами и не содержала эквивалентные (дублирующие) концепты. Для этого необходимо, чтобы отображения M_U, M_c, M_A, M_R на одинаковых концептах онтологий ИС совпадали. Результирующая онтология определяет соответствия концептов и правила их интерпретации между ИС, что позволяет успешно установить их взаимодействие.

Информационная система $U' =< \overline{O}, \overline{U}, \overline{M_U}, \overline{M_R} >$ называется *интегрированной* на множестве ИС $U^{\bar{O}}$, если $U^{\bar{O}} = \{U_1^{O^1}, U_2^{O^2}, \dots, U_N^{O^N}\}$ непротиворечиво, т.е. существуют $\overline{M_U}: \overline{U} \to \bar{C}, \overline{M_c}: \bar{C} \to 2^{\bar{A}}, \overline{M_A}: \bar{A} \to \bar{G}$, $M_R: \overline{U} \times \overline{U} \to \bar{R}$, являющиеся расширением соответствующих отображений: $M_{c_i}, M_{A_i}, M_{U_i}$ $(1 \le i \le N)$.

Для осуществления согласованного изменения данных в ИС необходимо установление между онтологиями семантических зависимостей, которые определяют семантическую близость концептов. Таким образом, цель интеграции заключается в сохранении соответствия множества онтологий информационных систем заданному набору семантических зависимостей.

Под семантической зависимостью, заданной на онтологии O, предполагается z-предикат, заданный на $\bar{O}$.

Множество семантических зависимостей $Z = \{z^1, z^2, z^3, z^4, z^5\}$ непротиворечиво, если существует онтология O, которая удовлетворяет зависимости z_i.

На практике зависимость между онтологиями необходимо сводить к зависимостям между концептами, которые в них входят. Они были рассмотрены, проанализированы и отнесены в следующие 5 классов:

1. Эквивалентность z^1: $map(C_1) = C_2, if\ (S(C_1, C_2)|\forall c_i \in O_1, \forall c_j \in O_2) > b$, где b – порог меры семантической близости $S(C_1, C_2)$, при которой строится отображение концепта C_1 в онтологию O_2.

2. Обобщение $(C_1 \overset{z^2}{\rightarrow} C_2)$, где отображение $z^2: C_1 \rightarrow C_2$- отображение, ставящее в соответствие концепту C_1 множество концептов C_2.

3. Уточнение $(C_1 \overset{z^3}{\rightarrow} C_2)$, где $z^3: C_1 \rightarrow C_2$- отображение, ставящее в соответствие множеству концептов C_1 концепт C_2.

4. Частичная эквивалентность z^4. $(C_1 \overset{z^4}{\rightarrow} C_2)$.

Пересечение множеств атрибутов концептов C_2 и C_1 $(A^2 \cap A^1 \neq \emptyset)$ свидетельствует о наличии общих атрибутов. Это означает, что существует некоторый концепт C, являющийся надклассом для концептов C_2 и C_1, а сами концепты принадлежат одному уровню иерархии.

5. Различие z^5. Пустое пересечение множеств атрибутов концептов C_2 и C_1 $(A^2 \cap A^1 = \emptyset)$.

Модель системы интеграции данных на основе онтологий представим в виде кортежа:

$$< O, U^O, Z, F, map >, \quad (8)$$

где $O =< C, A, G, D, M_A, M_c, R, I >$ – онтология ИС, U^O – информационная система с онтологией O, $Z = \{z^1, z^2, z^3, z^4, z^5\}$ – множество семантических зависимостей, $F: H^O \rightarrow H^O$ – такое отображение, что $\forall U^O \in H^O$, $\forall z \in Z$, выполнено $z(F(U^O))$, $map: O_i \rightarrow O_j$ – отображение онтологий.

Для численной оценки семантической близости концептов онтологий авторами выбран подход, основанный на результатах исследований профессора университета Мангейма (Германия) A. Maedche [4, 5]. В соответствии с этим рассматриваются атрибутивная, таксономическая и реляционная меры, результаты измерений с использованием каждой из них с учетом весовых коэффициентов и используются для комплексной оценки семантической близости.

При этом авторами предлагается определять атрибутивную меру не как пересечение диапазонов числовых значений атрибутов концептов, а как отношение пересечения множеств атрибутов к объединению множеств атрибутов концептов. Предлагается также определять весовые коэффициенты автоматически с использованием генетического алгоритма. Основные преимущества предлагаемого подхода заключаются в выявлении ключевых концептов для построения результирующей онтологии, устранения субъективности описаний понятий онтологии и зависимости от точек зрения разработчиков онтологий.

Определим $\mathrm{S^T}(\mathrm{c_i}, \mathrm{c_j})$ как мера близости двух концептов на основе их положения, $\mathrm{S^R}(\mathrm{c_i}, \mathrm{c_j})$ – мера близости двух концептов на основе сопоставления их отношений, $\mathrm{S^A}(\mathrm{c_i}, \mathrm{c_j})$ – мера близости двух концептов на основе сопоставления атрибутов и значений атрибутов концептов.

Мера близости $S(c_i, c_j)$ двух концептов c_i онтологии O и c_j онтологии O' определяется как:

$$S(c_i, c_j) = t \cdot S^T(c_i, c_j) + r \cdot S^R(c_i, c_j) + a \cdot S^A(c_i, c_j), \qquad (9)$$

где t – вес, определяющий важность меры близости $S^T(c_i, c_j)$; r – вес, определяющий важность меры близости $S^R(c_i, c_j)$; a – вес, определяющий важность меры близости $S^A(c_i, c_j)$.

С учетом того, что $t, r, a \in [0; 1], t + r + a = 1$, $S(c_i, c_j) \in [0; 1]$, причем если концепты идентичны $c_i = c_j$, тогда $S(c_i, c_j) = 1$, если концепты различны и не имеют общих характеристик, тогда $S(c_i, c_j) = 0$.

Для автоматического определения параметров t, r, a используется генетический алгоритм, где индивид представляется в виде тройки генов (t, r, a). В роли функции приспособленности выступает целевая функция:

$$f_{t,r,a} = t \cdot S^T(c_i, c_j) + r \cdot S^R(c_i, c_j) + a \cdot S^A(c_i, c_j).$$

К сформированной популяции потенциальных решений со следующими ограничениями $t, r, a \in [0; 1], t + r + a = 1$ применяются стандартные операторы отбора, кроссовера и мутации.

Критерий выбора: максимизация суммы мер семантической близости между концептами двух онтологий.

$$f_{t,r,a} = \sum_{\substack{c_i, c_j \in C \\ c_i \neq c_j}} S(c_i, c_j).$$

Для выделения меры семантической близости, при которой концепты эквивалентны, необходимо выбрать пороговое значение меры близости. Разработан метод определения

критерия подобия концептов для классификации отображений в пять групп: эквивалентность, частичная эквивалентность, обобщение, уточнение, неопределенность.

$$b = max\left(S\left(c_i, c_j\right)\middle|\forall c_i \in O_1, \forall c_j \in O_2\right) * (1 - p_1), \qquad (10)$$

где p_1 – процент, при котором b считается порогом подобия для определения эквивалентности концептов.

$$q = \min\left(S\left(c_i, c_j\right)\middle|\forall c_i \in O_1, \forall c_j \in O_2\right) * (1 - p_2), \qquad (10)$$

где p_2 – процент, при котором c считается порогом подобия для определения отсутствия эквивалентности концептов.

Рассмотренная математическая модель реализована на ЭВМ в рамках специального программного обеспечения, использованного при интеграции онтологий, построенных на объектных схемах информационных систем управления учебным процессом и финансового планирования вуза. Обе системы были разработаны независимо друг от друга в период, предшествовавший рассматриваемому исследованию, и функционировали на основе использования собственных локальных баз данных, обмен информацией между которыми осуществлялся с помощью программ-конвертеров.

В результате проведенного вычислительного эксперимента была создана интегрированная онтология, позволившая в короткие сроки объединить локальные базы данных упомянутых систем, исключить дублирование, а также обеспечить целостность и непротиворечивость представленных в них сведений.

Кроме того, аналогичная работа была проведена экспертом-аналитиком, соответствующие результаты представлены в таблице 1.

Таблица 1. Сравнение параметров процесса отображения онтологий

Способ интеграции	Найденные семантические зависимости				Критерий оценки (средние значения)		
	Обобщение	Уточнение	Эквивалентность	Частичная эквивалентность	Полнота (R)	Точность (P)	Мера (F_1)
Эксперт	7	3	4	14	0,86	0,82	0,86
Модель	12	3	8	16	0,98	0,94	0,98

Заключение

Построенная математическая модель интеграции онтологий ИС адекватно описывает их семантические особенности. Алгоритм интеграции с использованием онтологий в целом лишен многих недостатков, присущих чисто техническим методам, и предоставляет

возможность разработки интегрированных ИС, работающих с информацией на семантическом уровне. Практическое использование рассмотренных методов моделирования позволило в короткие сроки и с высоким качеством объединить локальные базы данных систем управления учебной деятельностью и финансового планирования в процессе развития ИАИС Бийского технологического института.

Список литературы

1. Бубарева О. А., Попов Ф. А., Ануфриева Н. Ю. Использование онтологий с целью интеграции данных в рамках автоматизированных информационных систем ВУЗов // Фундаментальные исследования. – 2011. – № 12 (часть 1). – С. 85-88.
2. Бубарева О. А., Попов Ф. А. Подсистема расчета себестоимости образовательной услуги в составе интегрированной автоматизированной информационной системы ВУЗа// Современные проблемы науки и образования. – 2011. – № 6; URL: www.science-education.ru/100-5053 (дата обращения: 16.03.2012).
3. Бездушный А. А. Математическая модель системы интеграции данных на основе онтологий // Журнал «Вестник НГУ», серия «Информационные технологии». – Новосибирск, 2008. – Т.6. Вып. 2. – С. 15-40.
4. Botzenhardt, A.; Maedche, A. & Wiesner, J.: Developing a Domain Ontology for Software Product Management. Proceedings of the 5th International Workshop on Software Product Management (IWSPM-2011), Trento, Italy. IEEE Xplore, Digital Library, 2011.
5. Maedche A., Zacharias V. // Proc. 6th European PKDD Conf. LNCS V. 2431. Berlin: Springer, 2002. P. 348.

Рецензенты:

1. Оскорбин Николай Михайлович, д.т.н., профессор, заведующий кафедрой теоретической кибернетики и прикладной математики ФГБОУ ВПО «Алтайский государственный университет».

2. Темербекова Альбина Алексеевна, доктор педагогических наук, профессор кафедры алгебры, геометрии и методики преподавания математики Горно-Алтайского государственного университета, зав. научно-исследовательской лаборатории «Инновационные образовательные технологии» ГАГУ.