УДК: 004.65

ИСПОЛЬЗОВАНИЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА В КОНТЕКСТЕ РЕШЕНИЯ ЗАДАЧИ НАХОЖДЕНИЯ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ЭЛЕМЕНТОВ НЕОДНОРОДНЫХ ОНТОЛОГИЙ

О.А. Бубарева, Ф.А. Попов

В статье рассматривается задача объединения онтологий информационных систем (ИС) с целью последующего установления взаимодействия схем ИС. 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 модифицированный генетический алгоритм. На основе предложенного метода реализована программная система интеграции данных информационных систем.

Ключевые слова: генетический алгоритм, онтология, интеграция

Введение

На сегодняшний день актуальным для организаций и учреждений является построение интегрированных автоматизированных информационных систем (ИАИС), обеспечивающих поддержку различных бизнес-процессов, а также возможность формирования агрегированной информации для принятия управленческих решений [2,5]. Получение интегрированной информации зависит от эффективного взаимодействия входящих в структуру ИАИС информационных систем (ИС) с различными стандартами описания и представления данных. В структуру ИАИС, как правило, входят гетерогенные ИС, имеющие собственные локальные модели данных [4]. При их слиянии в глобальную модель порождается ряд конфликтов, в частности: использование различных терминов для обозначения одних и тех же понятий; различного рода семантические конфликты. Анализ построения интегрированных систем показал, что в процессе их создания для решения проблемы неоднородности применяются методы, основанные на использовании единой онтологии верхнего уровня.

Как отмечает N. Guarino [7], любая ИС имеет свою онтологию, поскольку она приписывает значение каждому представленному в ней символу (имени), используемому в соответствии с присущим ей взглядом на мир. Для обеспечения совместной работы неоднородных ИС в контексте предметной области задачи необходимо согласовать онтологии, лежащие в их основе. Каждая онтология ИС, построенная разными группами экспертов, носит субъективный характер и обладает собственными категориями абстракций. Именно по этой причине объединение онтологии с целью последующего установления взаимодействия схем ИС является серьезной проблемой.

Нахождение семантической близости элементов неоднородных онтологий

Одним из вариантов решения такой проблемы является нахождение семантически близких элементов онтологий (концептов). 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.

Построение математической модели интеграции данных ИС с учетом сопоставления их онтологических спецификаций создает возможность для измерения близости (подобия) концептов онтологий [3].

Для численной оценки семантической близости концептов онтологий выбран подход, основанный на результатах исследований А.Ф. Тузовского и профессора университета Мангейма A. Maedche [6]. В соответствии с этим рассматриваются атрибутивная, таксономическая и реляционная меры с учетом весовых коэффициентов.

Данный метод был адаптирован для расчета семантической близости двух неоднородных онтологий. Модификация данного метода заключается в способе нахождения атрибутивной и таксономической составляющих, а также в применении генетического алгоритма для нахождения весовых коэффициентов. При этом предлагается определять таксономическую меру как отношение пересечения множеств терминов к объединению множеств терминов концептов. Основные преимущества предлагаемого подхода за-

ключаются в нахождении ключевых концептов, устранении субъективности их описаний и зависимости от точек зрения разработчиков онтологий.

Определим $S^T(c_i,c_j), S^R(c_i,c_j), S^A(c_i,c_j)$, соответственно, как меру близости двух концептов на основе их положения, сопоставления их отношений, а также сопоставления атрибутов и их значений.

Для оценки таксономической близости двух понятий $S^T(c_i,c_j)$ вводятся два показателя, основанные на сравнении множеств концептов $C_d(c_i)$:

$$S^T(c_i,c_j)=\begin{cases}1, если \quad c_i=c_j\\ \dfrac{\left|PL_p(c_i)\cap PL_p(c_j)\right|}{\left|PL_p(c_i)\cup PL_p(c_j)\right|}, если \quad c_i\neq c_j\end{cases}, \quad (1)$$

где $PL_p(c_i)=\{L_i\in L \mid P_c(c_i)=L_i\}$

- множество терминов концепта c_i.

Для оценки реляционной близости предполагается, что если два концепта имеют одинаковые отношения R_1 (таксономические) с третьим концептом, то они более похожи, чем два концепта, которые имеют разные отношения R_1.

Предположим, что

$C_r(c_i)=\{c_j\in C \mid R_1(c_i,c_j)\vee R_2(c_i,c_j)\vee R_3(c_i,c_j)\vee c_j=c_i\}$

- множество, содержащее концепты, у которых существуют отношения R_1,R_2,R_3.

$R_1^{Tr}=\{(c_i,c_j):(\exists c_i^1...c_i^n\in C: R_1(c_i,c_i^1)...R_1(c_i^n,c_j))\}$,

$R_{Ei}(c_i)=\{R_i: R_i\in R\wedge((c_i,C_r(c_i))\in R_1^{Tr}\}$ - множество отношений концепта c_i, которые определяют концепты из множества C_r.

$R_E(c_i,c_j)=R_{Ei}(c_i)\cap R_E(c_j)$ - определяем общие отношения $R_{Ei}(c_i)$ концептов c_i и c_j.

Определим отношение ассоциативности концептов как:

$$R_A(R,c_j)=\{c_i: c_i\in C\wedge R(c_j,c_i)\} \quad (2)$$

Реляционная мера близости $S^R(c_i,c_j)$ концептов c_i и c_j позволяет оценить схожесть двух концептов, исходя из их отношений с другими концептами.

$$S^R(c_i,c_j)=\frac{\sum_{R\in R_E}\sum_{r1}\max[S(r1,r2)\mid r2\in R_A(R,c_j)]}{\left|R_E\right|}, \quad (3)$$

где $r1\in R_A(R,c_j)$.

Сравним атрибуты двух концептов.

Зададим множество атрибутов, принадлежащих концепту c_i:

$A^{C_i}=\{A_k^{C_i}, k\in[1..n_1]\}$, где n_1 - количество атрибутов концепта c_i.

$A^{C_j}=\{A_k^{C_j}, k\in[1..n_2]\}$,

где n_2 -количество атрибутов концепта c_j.

Атрибутивная мера близости $S^A(c_i,c_j)$ концептов c_i и c_j определяется соответствием их атрибутов: общих $A^{C_i}\cap A^{C_j}$, различных $A^{C_i}\setminus(A^{C_i}\cap A^{C_j})$, то есть c_i, которых нет в c_j и атрибутов c_j, отсутствующих у c_i : $A^{C_j}(A^{C_i}\cap A^{C_j})$.

Атрибутивная мера близости $S^A(c_i,c_j)$ удовлетворяет аксиомам монотонности, независимости, разрешимости и инвариантности и определяется формулой:

$$S^A(c_i,c_j)=\frac{\left|A^{C_i}\cap A^{C_j}\right|}{\left|(A^{C_i}\setminus(A^{C_i}\cap A^{C_j})\cup A^{C_j}(A^{C_i}\cap A^{C_j}))\right|}, \quad (4)$$

где A^{C_i} - множество атрибутов концепта c_i, A^{Cj} - множество атрибутов концепта c_j.

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

$$S(c_i,c_j)=t\cdot S^T(c_i,c_j)+r\cdot S^R(c_i,c_j)+a\cdot S^A(c_i,c_j), \quad (5)$$

где t,r,a - коэффициенты, определяющие важность мер близости:

$S^T(c_i,c_j), S^R(c_i,c_j), S^A(c_i,c_j)$.

$t,r,a\in[0,1], t+r+a=1;\ S(c_i,c_j)\in[0,1]$.

$$\begin{cases}S(c_i,c_j)=1, концепты\ одинаковые,\\ S(c_i,c_j)=0, концепты\ различны.\end{cases}$$

Для решения задачи нахождения весовых коэффициентов предлагается использование генетического алгоритма, который обеспечивает поиск решения для функций, имеющих несколько экстремумов. На рисунке 1 представлен модифицированный генетический алгоритм. Для задач подобного рода это

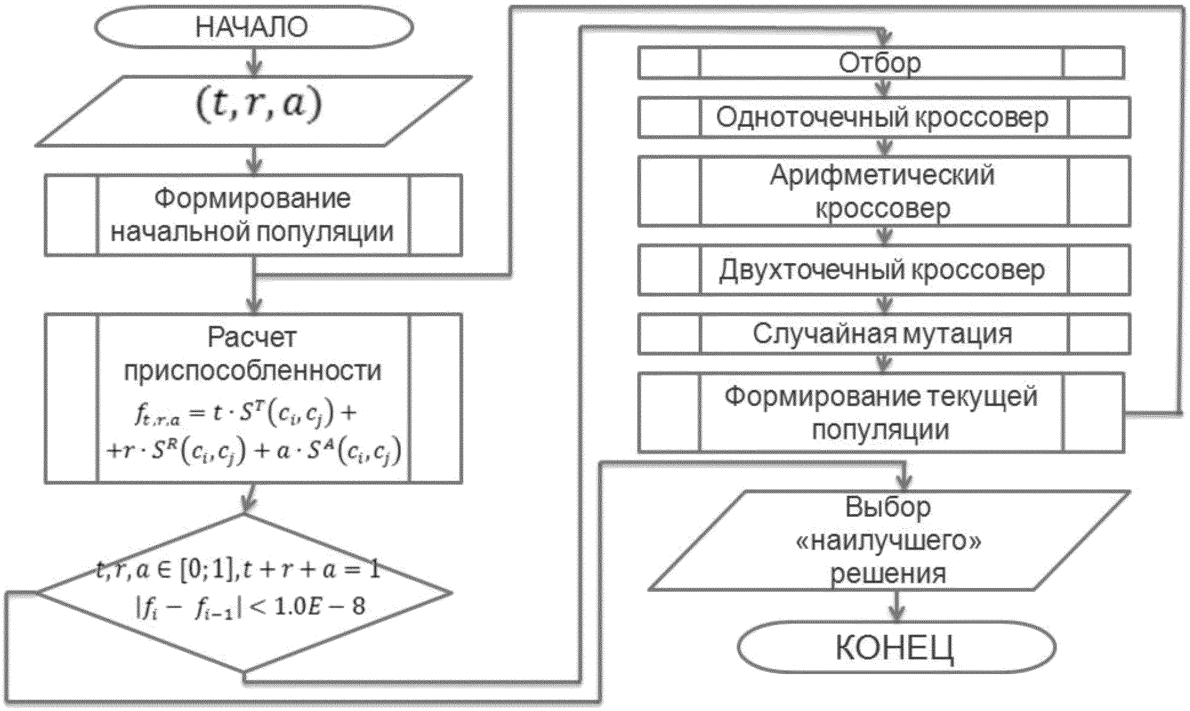

Рисунок 1 – Блок - схема генетического алгоритма

один из наиболее распространенных и эффективных методов решения.

Задача по оценке семантической близости концептов онтологии относится к группе задач оптимизации с ограничениями. В общем виде ее можно представить следующим образом:

$$\max f_{t,r,a}(\bar{x}) \bar{x} = (t,r,a) \in F \subseteq S$$

$$t,r,a \in [0;1] \quad t + r + a = 1,$$

где $\bar{x}$ – вектор решений, удовлетворяющий всем ограничениям и называемый допустимым решением, F – область допустимых решений, S – вся область поиска. Задачу оптимизации можно сформулировать следующим образом: найти $\bar{x}' \in F$ такой, что

$$f_{t,r,a}(\bar{x}') \geq f_{t,r,a}(\bar{x}) \forall \bar{x} \in F .$$

Для решения данной задачи конструируется хромосома, которая состоит из набора генов (t,r,a). Начальная популяция задается случайно сгенерированным набором значений. Каждое новое поколение генерируется при помощи оператора кроссинговера. Родительские пары выбираются методом турнирного отбора. В каждой новой хромосоме некорректные гены подвергаются случайной мутации. Все повторяющиеся хромосомы из популяции удаляются.

В роли функции приспособленности выступает целевая функция (5).

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

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

В результате проведенного исследования были определены наиболее эффективные генетические операторы и параметры. Анализ результатов вычислительных экспериментов показал, что генетический алгоритм выдает лучший результат при использовании ряда операторов кроссинговера, отбора и мутации. В генетический алгоритм были включены следующие генетические операторы: отбор, 30% одноточечный кроссинговер, 40% арифметический кроссинговер и 30% двухточечный; случайная мутация.

Использование ряда генетических операторов, выявленных в эксперименте, позволяет получить поколение особей с наилучшим значением целевой функции и приводит общему сокращению времени решения задачи.

Оценка достоверности результатов генетического алгоритма проводилась для случая нахождения концептов "Частично эквивалентны" по методу, описанному в работе А.А. Асанова [1]. Для этого введены коэффициен-

ты абсолютной ошибки E_{abs} и относительной ошибки E_{rel}.

Число исходных частично эквивалентных концептов равно 57. Число найденных позиций равно 52. Тогда коэффициент абсолютной ошибки будет равен 5, коэффициент относительной ошибки равен 0,17.

Степень покрытия cd множеством частично эквивалентных концептов множества исходных для множества найденных концептов она равна $1 - E_{rel} = 0.83$.

Таким образом, полученное значение степени покрытия cd показывает, что достоверность найденных частично эквивалентных концептов достаточно высока.

Проведен сравнительный анализ с методом перебора и методом градиентного спуска. При использовании метода перебора с увеличением числа концептов в онтологии увеличивается количество вариантов решений. При методе градиентного спуска выбираются некоторые случайные значения параметров, а затем, изменяя эти значения, добиваются наибольшей скорости роста целевой функции. Достигнув локального максимума, такой алгоритм останавливается, и поэтому для поиска глобального оптимума потребуются дополнительные усилия. Такой метод не гарантирует оптимальности найденного решения. В результате анализа было выявлено, что предложенный генетический алгоритм обладает ускоренной сходимостью и показывает наилучший конечный результат.

Метод вычисления семантической близости концептов позволяет количественно оценить сходство между понятиями. Для каждого концепта одной онтологии формируется множество релевантных семантических концептов другой онтологии. С целью ранжирования элементов результирующего множества необходимо определить пороговые значения меры близости.

Заключение

На основе предложенного метода реализована программная система интеграции данных информационных систем. 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.

В результате проведенных экспериментов было построено отображение онтологий, позволившее в короткие сроки объединить локальные базы данных упомянутых систем, исключить дублирование, а также обеспечить целостность и непротиворечивость представленных в них сведений. После интеграции онтологии ИС становится возможным интерпретировать информацию из одной ИС средствами другой ИС. С помощью этой программы были интегрированы данные ИС управления учебным процессом с данными ИС финансового планирования.

СПИСОК ЛИТЕРАТУРЫ

1. Асанов, А.А. Исследовано в России [Электронный ресурс]. – Режим доступа: http://zhurnal.ape.relarn.ru/.
2. Бубарева, О. А. Использование онтологий с целью интеграции данных в рамках автоматизированных информационных систем ВУЗов /О.А. Бубарева, Ф.А. Попов, Н.Ю. Ануфриева // Фундаментальные исследования. – 2011. – № 12 (часть 1). – С. 85-88.
3. Бубарева, О.А., Математическая модель процесса интеграции информационных систем на основе онтологий // О.А. Бубарева, Ф.А. Попов // Современные проблемы науки и образования. – 2012. – № 2; URL: www.science-education.ru/ 102-6030 (дата обращения: 21.02.2013).
4. Попов, Ф.А. Подходы к интеграции научно-производственных и образовательных информационных ресурсов / Ф.А. Попов // Ползуновский вестник. – 2004. − №3. − С. 19-23.
5. Попов, Ф.А. Проблемы проектирования баз данных применительно к информационно-управляющим системам для научно-производственного объединения / Ф.А. Попов // Ползуновский вестник. – 2006. − №2-2. С. 127-133.
6. Maedche A., Zacharias V. // Proc. 6th European PKDD Conf. LNCS V. 2431. Berlin: Springer, 2002. P. 348.
7. N. Guarino, Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, P. 3-15.

Соискатель ***Бубарева О.А.*** *– angel@bti.secan.ru; д.т.н. проф.* ***Попов Ф.А.*** *– pfa@bti.secna.ru – Бийский технологический институт тел.(385-4)-43-53-00*