05.13.17

ИССЛЕДОВАНИЕ МЕХАНИЗМОВ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЙ НАД МНОЖЕСТВОМ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

О.А. Бубарева

Бийский технологический институт АлтГТУ, г. Бийск

Актуальной проблемой интеграции ИС является анализ, согласование и отображение их онтологий, которые построены разными способами. В статье приводится краткий обзор процесса автоматического построения первоначальных версий онтологий предметной области. Предложен подход к семантической интеграции неоднородных онтологий сложных информационных систем. Основная идея заключается в рассмотрении онтологий из разных предметных областей и построенных разными способами. В статье описан метод вычисления семантической близости концептов, который позволяет количественно оценить сходство между понятиями. Соответствия между элементами (концептами) онтологий разделяются на несколько составляющих: лексическую, атрибутивную и реляционную. Для каждого концепта одной онтологии формируется множество релевантных семантических концептов другой онтологии. С целью ранжирования элементов результирующего множества предлагается определять пороговые значения меры близости. В работе предложен метод классификации значений близости концептов для установления их корректного отображения. Следует отметить, что основное ядро онтологии адекватно по семантики и повторяет построенную экспертами модель. Высокая точность результатов объясняется применением расширенного набора вариантов взаимного позиционирования концептов. Модель процесса интеграции применима к широкому кругу предметных областей и не сложна в реализации.

Ключевые слова: онтологии, отображение онтологий, архитектура системы, интеграция данных,генетический алгоритм, семантическая близость.

ВВЕДЕНИЕ

С целью реализации электронного университета в ВУЗе должно функционировать единое интегрированное информационное пространство, которое включает множество распределенных информационных систем (ИС). Усложнение функций, реализуемых динамическими распределенными ИС, приводит к увеличению трудоемкости разработки и сопровождения таких систем. Интегрированные распределенные ИС могут состоять из разнородных моделей данных - онтологий, а также алгоритмов их обработки [1-7]. Первоначальные версии онтологий ИС могут быть построены вручную различными группами экспертов или автоматически.

Актуальной проблемой интеграции ИС является анализ, согласование и отображение их онтологий, которые построены разными способами.

В данной статье рассматриваются популярные подходы к автоматизированному построению первоначальных версий онтологий ИС.

ПОСТРОЕНИЕ ОНТОЛОГИИ

В научных исследованиях [8] для решения вопроса автоматического построения онтологии предлагается подход, основанный на лексико-синтаксическом шаблоне (LSPL) для коллекций русскоязычных текстов с использованием коммуникативных грамматик. Такие лингвистические конструкции способны показать семантические связи между терминами и могут быть применимы для определения концептов и отношений между ними при построении онтологии по тексту на естественном языке. Применение коммуникативных грамматик позволяет анализировать смысл синтаксических конструкций текста, а также выделять синтаксемы – минимальные синтактико-семантические единицы языка.

В работе [8] сделана попытка усовершенствовать уже существующие LSPL-шаблоны за счет введения в них нового компонента – категориальный смысл синтаксем. В данной работе неучтен тот факт, что коммуникативные грамматики составляются вручную, и это затрудняет их использование при идентификации множества разнотипных концептов онтологии.

В научных работах Найхановой Л.В. в качестве метода генерации грамматик используется генетический алгоритм и механизм автоматного программирования [9]. Однако анализ таких работ показывает, что данные методы пока подходят лишь для узкого круга задач.

Во многих научных исследованиях для сравнения онтологий предполагается использование тезауруса, а для выявления схожих слов используются лексические отношения – синонимия, гипонимия, омонимия.

В работе [10] сформулирован метод автоматического построения и сравнения контекстов понятий различных онтологий для оценки их семантической близости в процессе онтологической интеграции. Метод позволяет устранить субъективности неформальных описаний элементов онтологии и исключает необходимость использования специализированных тезаурусов. Отдельное внимание в работе направлено на создание процедур комплексного анализа корпуса текстов и разработку алгоритмов формирования и сравнения контекстов онтологий.

В работе [11] процесс автоматической интеграции онтологий по набору текстовых документов включает функции извлечения объектов предметной области (концептов), семантических отношений и регулярных выражений на основе генетических алгоритмов. Методы извлечения семантических отношений основаны на базе лексико-синтаксических шаблонов. А суть метода идентификации объектов заключается в извлечении цепочек символов и их соотношения с терминами тезауруса по той или иной известной семантической категории.

На сегодняшний день в научных работах сформулировано множество методов по отображению информационных моделей данных, описаны различные способы интеграции, которые в основном ориентированы на конкретные предметные области [12]. Исследований по автоматическому построению единой онтологий при отображении произвольных онтологических моделей немного и нет четкого понимания структуры процесса интеграции онтологий. Обычно они базируются на методах объектных или реляционных схем ИС. Так как каждая онтология субъективна и обладает собственными категориями абстракций, то процесс их интеграции является достаточно трудоемким и обычно осуществляется на основе принятых экспертом решений, что может приводить к долгим философским спорам между участниками. Поэтому автоматическая интеграция онтологий с минимальным участием высококвалифицированных специалистов-экспертов для динамически развивающихся систем является актуальной задачей.

ПОСТАНОВКА ЗАДАЧИ

Под неоднородностью онтологий подразумевается, что одна и та же предметная область может быть описана онтологиями по-разному (рис. 1). Причиной этому может быть особенности подходов к описанию спецификаций понятий. Поэтому семантика одного понятия в разных онтологиях может быть сходной.

Задача настоящего исследования заключается в разработке модели интеграции множества неоднородных онтологий сложных динамических ИС, которая сводится к интеллектуальной обработке данных для автоматизированного построения отображений онтологических спецификаций с учетом согласования на уровнях модельной и понятийной семантики.

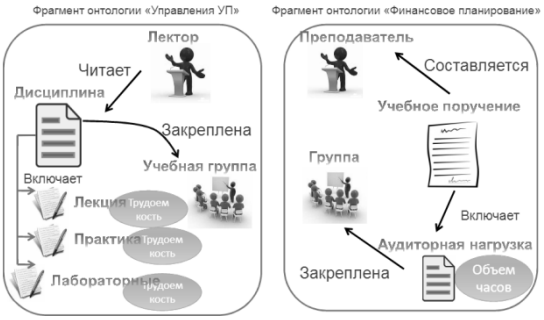

Рис. 1 – Пример онтологии учебного процесса

В качестве решения этой задачи, предлагается разработать алгоритмы автоматического извлечения данных из слабоструктурированных источников, методы на основе онтологического подхода для структурирования и интерпретации знаний, а также комплекс программ для анализа данных.

ПОДХОД К ОТОБРАЖЕНИЮ ОНТОЛОГИЙ

Онтологии носят субъективный характер, так как создаются разными рабочими группами, но с точки зрения своей ИС и решаемых задач они корректны. Но в процессе интеграции возникают проблемы сравнения разных онтологий ПО, которые заключаются в различии имен понятий, отношений и атрибутов, в разбиении предметной области на понятия.

С целью обеспечения согласованного изменения моделей данных в ИС необходимо решить задачу по нахождению сходств и различий в концептах онтологий. При этом рассчитывается семантическая близость и устанавливаются зависимости между концептами онтологий. Таким образом, цель интеграции заключается в нахождении для каждого концепта одной онтологии подобного концепта другой онтологии, а также семантических зависимостей между концептами двух онтологий ИС.

Математическая модель системы интеграции онтологий ИС представлена в виде кортежа

$$S = \left\langle O, U^O, Z, map \right\rangle, \qquad (1)$$

где O - множество онтологий ИС;

Z – множество семантических зависимостей между концептами;

U^O – информационная система;

$map : O_i \to O_j$ – отображение концептов онтологий.

Онтология информационной системы представлена в следующем виде:

$$O = \langle C, A, L, P_A, P_C, R \rangle, \quad (2)$$

C - множество концептов онтологии;

A - множество атрибутов концептов;

L - словарь, в котором определяются профессиональные термины организации;

$P_A : C \to 2^A$ – отображение, задающее для каждого концепта множество его атрибутов;

$P_C : C \to 2^L$ – функция интерпретации концептов, сопоставляет концепту набор терминов из словаря L;

R – множество отношений между концептами.

Мера семантической близости является аддитивной сверткой трех составляющих по разным типам характеристик концептов с учетом весовых коэффициентов. Она имеет область определения от 0 до 1. Было предложено изменение способа нахождения ряда составляющих, что позволило расширить область применения данного метода на сопоставление концептов из разных онтологий. Для нахождения весовых коэффициентов, определяющих важность той или иной составляющей, применяется генетический алгоритм [1].

Для определения лексической составляющей сравниваются множества синонимичных терминов словаря. Для определения атрибутивной составляющей находятся схожие атрибуты путем точного сопоставления или редакторского расстояния. Для определения реляционной составляющей сравниваются множества концептов, у которых есть отношения с исходным концептом.

На основе использования математической модели разработан алгоритм интеграции онтологий ИС, состоящий из несколько этапов.

Этап 1. Формирование множества допустимых вариантов сопоставления концептов онтологий.

Шаг 1. Определение множества концептов. Формируется множество концептов исходных онтологии, которые необходимо интегрировать. Определяются связные концепты, которые следует учитывать при формировании допустимых вариантов сравнения онтологий. Формируются множество атрибутов концептов и ограничений, предъявляемых к онтологиям.

Шаг 2. Сопоставление концептов. Происходит сопоставление концептов двух онтологий. Формируется расширенное множество, состоящее из вариантов «концепт-концепт». Строится дерево решений, учитывающее взаимное расположение концептов.

Шаг 3. Определение значения семантической близости концептов. Каждый вариант «концепт-концепт» получает оценку следующих составляющих меры семантической близости: лексической, реляционной и атрибутивной. Формируется множество Парето, в которое включены варианты «концепт-концепт», имеющие наилучшую оценку хотя бы по одной из составляющих меры близости. Так как требования противоречивы, то при построении множества нужно использовать интервальные оценки или нечеткие градации, в противном случае данное множество может оказаться пустым. Согласно полученным оценкам проводится ранжирование вариантов в порядковой шкале. Формируется множество допустимых вариантов возможного отображения концептов, включающее разнотипные решения, удовлетворяющие требованиям онтологий.

Этап 2. Выбор наилучшего варианта отображения концептов.

Шаг 1. Расчет весовых коэффициентов для меры семантической близости концептов. Расчет весовых коэффициентов предусматривает также определение критериев, позволяющих оценить пригодность вариантов для построения отображения концептов. При этом следует принять во внимание такие ситуации, когда концепты частично эквиваленты или один концепт является уточнением или обобщением другого концепта. Это непосредственно влияет на функцию выбора варианта позиционирования каждого концепта.

Шаг 2. Оценивание вариантов отображения концептов. В соответствии с полученными оценками семантической близости для каждого варианта из допустимого множества выбирается модель взаимного позиционирования концептов (эквивалентность, частичная эквивалентность, уточнение, обобщение, неопределенность).

Этап 3. Выполнение интеграции и оценка результатов.

Шаг 1. Устанавливается отображение концептов. Концепту одной онтологии ставится в соответствие концепт из другой онтологии.

Шаг 2. Оценка результатов. Оцениваются правильность построенной результирующей онтологии. Оценивается эффективность работы программы по сопоставлению онтологий с помощью экспертной группы. Сопоставляются полученные результаты с планируемыми.

Шаг 3. Интеграция информационных систем. Согласно найденным связям между двумя онтологиями выполняется установление отображения реляционных моделей информационных систем. После этого происходит генерация SPARQL запросов, а также консолидация данных одной ИС в другую.

Согласно выбранной модели взаимного позиционирования концептов выполняются следующие операции:

1. При эквивалентности концептов происходит объединение атрибутов концептов.

2. Если один концепт позиционируется как «обобщение» другого концепта, то такие объекты

представляются как «класс-подкласс» соответственно. При этом идентичные атрибуты удаляются из подкласса.

3. Если один концепт рассматривается как «уточнение» другого концепта, то такие объекты представляться как «подкласс-класс» соответственно. В этом случае идентичные атрибуты удаляются и учитываются все существующие отношения этих концептов.

4. Если два концепта позиционируются как «частично эквивалентны», то создается новый концепт «надкласс», являющийся обобщением, при этом идентичные атрибуты удаляются из подкласса.

КЛАССИФИКАЦИЯ УРОВНЕЙ БЛИЗОСТИ КОНЦЕПТОВ

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.

Разработан метод классификации значений близости концептов для установления их корректного отображения (рис. 2).

Рассматривается вопрос поиска минимального порога b семантической близости $M(c_i, c_j)$ концептов c_i и c_j, при которой концепты принимаются эквивалентными.

$$b = \max\left(M(c_i, c_j) \middle| \forall c_i \in O_1, \forall c_j \in O_2\right) \times (p_1 / 100), \quad (3)$$

где p_1 – процент, при котором b принимается порогом подобия для установления эквивалентности и корректного отображения c_i и c_j.

Показано, что b – минимальный порог, при котором уменьшение этого значения приводит к невозможности полного отображения элементов онтологий.

Находится пороговое значение, при котором концепты принимаются частично эквивалентными.

$$q = \max\left(M(c_i, c_j) \middle| \forall c_i \in O_1, \forall c_j \in O_2\right) \times (p_2 / 100),$$

где p2 – процент, при котором q принимается порогом подобия для установления частичной эквивалентности концептов.

Показано, что q – минимальное значение в том смысле, что уменьшение этого значения приводит к некорректному отображению элементов онтологии.

Концепты принимаются различными, если имеют значение меры семантической близости не превосходящее порог q.

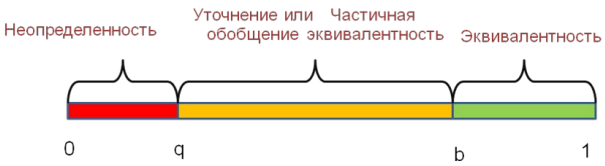

Рис. 2 – Метод классификации уровней семантического подобия концептов

Таким образом, можно построить модель единого интегрированного информационного пространства на основе онтологий информационных систем разных предметных областей. Это будет унифицированная точка входа информации из систем и источников данных в единое информационное пространство.

Построенная модель единого интегрированного информационного пространства наилучшим образом отражает ИАИС и служит основой для определения семантических зависимостей, а также дает возможность применить технологию интеграции данных информационных систем разных предметных областей.

Результатом математического моделирования является построение модели интеграции ИС, а также доказательство ее соответствия поставленной цели исследования. Применимость модели исследовалась при интеграции систем разных предметных областей вуза. Согласно проведенному анализу полученных результатов построенная модель интеграции ИС способна адекватно описывать исходную ситуацию. 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.

Для оценки качества работы были выбраны показатели точности, полноты и среднее значение F-меры. В итоге средний выигрыш алгоритма по качеству сопоставления составил примерно 40% (рис. 3).

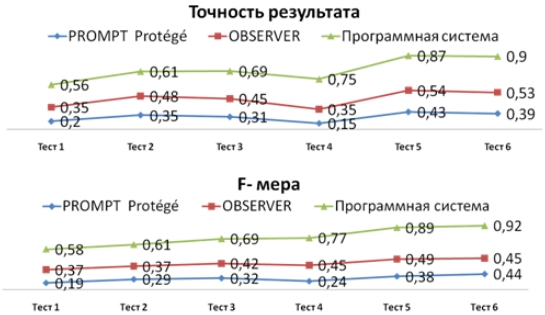

Рис. 3 – Сравнение результатов исследования

Полученные результаты позволяют считать успешной выполненную работу по разрешению семантических конфликтов при построении

онтологии неоднородных пересекающихся предметных областей.

ЗАКЛЮЧЕНИЕ

В статье представлен подход к интеграции онтологий информационных систем с распределенной архитектурой. Следует отметить, что основное ядро онтологии адекватно по семантики повторяет построенную экспертами модель. Высокая точность результатов объясняется применением расширенного набора вариантов взаимного позиционирования концептов. Модель процесса интеграции применима к широкому кругу предметных областей и не сложна в реализации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Olesya A. Bubareva. Ontology Integration in Complex Information Systems with Distributed Architecture // 19th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices EDM`2018: Conference proceedings. -Novosibirsk: NSTU Publishing polygraph center, 2018. -P. 216-219.

2. Ведриганов С.А. Оценка качества программных систем при связывании объектных спецификаций по семантике онтологического уровня /Ведриганов С.А., Бубарева О.А// Материалы конференции ИАМП. – Бийск. 2017. с. 35-37.

3. Бубарева О.А. Надежность интегрированных информационных систем. Информация и образование: границы коммуникаций. 2016. № 8 (16). – С. 79-81.

4. Попов Ф.А., Ануфриева Н.Ю., Бубарева О.А., Тютякин А.А. Информационная система управления финансами вуза // Материалы конференции: Фундаментальные науки и образование. 2012. – С. 176-179.

5. Жданов И.Р., Бубарева О.А. Решение задачи интеграции неоднородных баз данных в системе автоматизации заполнения индивидуального плана преподавателя // Материалы конференции: Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования. 2017. – С. 705-707.

6. Бубарева О.А. К вопросу разрешения семантических конфликтов при интеграции информационных систем // Материалы конференции. – Стерлитамак: АМИ,2018. – С. 44-46.

7. Бубарева О.А. Оценка качества информационных систем с распределенной архитектурой // Материалы конференции ИАМП. – Бийск. 2017. с. 32-34.

8. Романов С.В. О возможностях использования коммуникативных грамматик и LSPL-шаблонов для автоматического построения онтологий / Романов С.В., Сытник А.А., Шульга Т.Э. // Известия Самарского научного центра РАН. 2015. №2-5. URL: https://cyberleninka.ru/article/n/o-vozmozhnostyah-ispolzovaniya-kommunikativnyh-grammatik-i-lspl-shablonov-dlya-avtomaticheskogo-postroeniya-ontologiy (дата обращения: 23.07.2018).

9. Найханова Л. В. Модель генератора конечных преобразователей, основанная на применении генетического и автоматного программирования // Вестник МГУЛ – Лесной вестник. 2008. №6. URL: https://cyberleninka.ru/article/n/model-generatora-konechnyh-preobrazovateley-osnovannaya-na-primenenii-geneticheskogo-i-avtomatnogo-programmirovaniya (дата обращения: 24.07.2018).

10. Маслобоев А.В. Метод автоматического построения и сравнения контекстов понятий онтологий для оценки их семантической близости / Маслобоев А.В., Ломов П.А., Мавренков Н.М.// Труды Кольского научного центра РАН. 2010. №3. URL: https://cyberleninka.ru/article/n/metod-avtomaticheskogo-postroeniya-i-sravneniya-kontekstov-ponyatiy-ontologiy-dlya-otsenki-ih-semanticheskoy-blizosti (дата обращения: 23.07.2018).

11. Платонов А.В. Методы автоматического построения онтологий / А.В. Платонов, Е.А. Полещук // Программные продукты и системы. 2016. №2 (114). URL: https://cyberleninka.ru/article/n/metody-avtomaticheskogo-postroeniya-ontologiy (дата обращения: 24.07.2018).

12. Бубарева О.А. Методология оценки эффективности программных систем на всех этапах жизненного цикла // Материалы конференции. – Стерлитамак: АМИ, 2018. - С. 19-22.

Бубарева Олеся Александровна – к.т.н., доцент кафедры МСИА, Бийский технологический институт (филиал) ФГБОУ ВПО АлтГТУ, тел. (3854)435300, e-mail: angel@bti.secna.ru.

RESEACH OF MECHANISMS OF AUTOMATIC CONSTRUCTION OF ONTOLOGIES OVER MULTIPLE UNSTRUCTURED DATA

O.A. Bubareva
Biysk Technological Institute, Biysk

Abstract: The actual problem of integrating IP is analysis, matching and mapping of their ontologies, which are constructed in different ways. The article provides a brief overview of the process of automatic construction of the initial versions of domain ontologies. An approach to the semantic integration of heterogeneous ontologies of complex information systems is proposed. The basic idea is to consider ontologies from different subject areas, and constructed in different ways. The article describes a method for calculating the semantic proximity of concepts, which allows one to quantify the similarity between concepts. The correspondences between the elements (concepts) of ontologies are divided into several components: lexical, attributive and relational. For each concept of one ontology, a set of relevant semantic concepts of another ontology is formed. For the purpose of ranking the elements of the resulting set, it is proposed to determine the threshold values of the proximity measure. A method for classifying the levels of proximity of concepts to establish their correct mapping is proposed. It should be noted that, excluding some error in comparing ontology concepts with the help of the software system, the main ontology core is adequately based on semantics and repeats the model constructed by experts. The high accuracy of the results is due to the use. The model of the integration process is applicable to a wide range of subject areas and is not difficult to implement.

Index terms: ontology, ontology mapping, system architecture, data integration, genetic algorithm, semantic proximity.

REFERENCES

1. Olesya A. Bubareva. Ontology Integration in Complex Information Systems with Distributed Architecture // The International Conference of Young Specialists on Micro / Nanotechnologies and Electron Devices EDM`2018: Conference proceedings. -Novosibirsk: NSTU Publishing polygraph center, 2018. -P. 216-219.

2. S.Vedriganov. Evaluation of the quality of software systems when binding object specifications on the semantics of the ontological level / Vedriganov SA, Bubareva OA // Proceedings of the IAPS conference. - Biysk. 2017. p. 35-37.

3. Bubareva O.A. Reliability of integrated information systems. Information and education: the boundaries of communications. 2016. No. 8 (16). - P. 79-81.

4. Popov FA, Anufrieva N.Yu., Bubareva OA, Tyutyakin AA Information system of university financial management // Proceedings of the conference: Fundamental sciences and education. 2012. - P. 176-179.

5. Zhdanov IR, Bubareva OA Solution of the problem of integration of heterogeneous databases in the automation system for filling an individual teacher's plan // Conference materials: Lomonosov's readings in the Altai: fundamental problems of science and education. 2017. - P. 705-707.

6. Bubareva OA To the issue of resolving semantic conflicts in the integration of information systems // Proceedings of the conference. - Sterlitamak: AMI, 2018. - P. 44-46.

7. Bubareva O.A. Evaluation of the quality of information systems with distributed architecture. Proceedings of the IAPS conference. - Biysk. 2017. p. 32-34.

8. Romanov S.V. On the possibilities of using communicative grammars and LSPL-patterns for automatic ontology building / Romanov SV, Sytnik AA, Shulga TE // Proceedings of the Samara Scientific Center of the Russian Academy of Sciences. 2015. №2-5. URL: https://cyberleninka.ru/article/n/o-vozmozhnostyah-ispolzovaniya-kommunikativnyh-grammatik-i-lspl-shablonov-dlya-avtomaticheskogo-postroeniya-ontologiy (date of circulation: July 23, 2013).

9. Nayhanova, LV, "The model of the generator of finite converters, based on the use of genetic and automatic programming." Vestnik MSGL - Lesnoy Vestnik. 2008. № 6. URL: https://cyberleninka.ru/article/n/model-generatora-konechnyh-preobrazovateley-osnovannaya-na-primenenii-geneticheskogo-o-avtomatnogo-programmirovaniya (reference date: July 24, 2013).

10. Masloboev A.V. The method of automatic construction and comparison of contexts of ontologies to assess their semantic closeness / Masloboev AV, Lomov PA, Mavrenkov NM // Proceedings of the Kola Science Center of the Russian Academy of Sciences. 2010. № 3. URL: https://cyberleninka.ru/article/n/metod-avtomaticheskogo-postroeniya-i-sravneniya-kontekstov-ponyatiy-ontologiy-dlya-otsenki-ih-semanticheskoy-blizosti (application date: July 23, 2013).

11. Platonov A.V. Methods of automatic construction of ontologies / A.V. Platonov, E.A. Poleshchuk // Software products and systems. 2016. №2 (114). URL: https://cyberleninka.ru/article/n/metody-avtomaticheskogo-postroeniya-ontologiy (date of circulation: July 24, 2013).

12. Bubareva OA Methodology for evaluating the effectiveness of software systems at all stages of the life cycle // Proceedings of the conference. - Sterlitamak: AMI, 2018. - P. 19-22.

Bubareva Olesya Aleksandrovna – associate professor at the char of methods and means of measurement and automation, Biysk Technological Institute, (3854)435300, e-mail: angel@bti.secna.ru.