УДК 004.896

И. Б. Бондаренко, А. И. Иванов

ОРГАНИЗАЦИОННАЯ МОДЕЛЬ МНОГОАГЕНТНОЙ СИСТЕМЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ РАСПРЕДЕЛЕННЫХ ГЕТЕРОГЕННЫХ БАЗ ДАННЫХ СИСТЕМ АВТОМАТИЗИРОВАННОГО ПРОЕКТИРОВАНИЯ

Аннотация.

Актуальность и цели. Объектом исследования являются распределенные базы данных систем автоматизированного проектирования, имеющих различную структуру данных. Предметом исследования является процесс извлечения знаний из таких баз данных. Цель – разработка архитектуры подсистемы извлечения знаний из распределенных гетерогенных баз данных.

Материалы и методы. Распределенность источников данных, гетерогенность представленных в них данных и вычислительная сложность анализа данных большого объема обусловливают применение агентно-ориентированного подхода к достижению поставленной цели.

Результаты. Разработана организационная модель многоагентной системы извлечения знаний из распределенных гетерогенных баз данных. Описаны основные модели ролей агентов и их взаимодействие между собой.

Выводы. Основная часть архитектуры подсистемы извлечения знаний из распределенных гетерогенных источников определяется подсистемой подготовки набора данных и подсистемой интеллектуального анализа данных. Основные проблемы при разработке подсистем такого класса обусловлены различной структурой данных, представленных в локальных источниках, а также различной точностью, надежностью и полнотой данных.

Ключевые слова: база данных, интеллектуальный анализ данных, САПР, многоагентная система, извлечение знаний, слияние данных.

I. B. Bondarenko, A. I. Ivanov

AN ORGANIZATIONAL MODEL OF A MULTI-AGENT KNOWLEDGE DISCOVERY SYSTEM FROM DISTRIBUTED HETEROGENEOUS CAD-DATABASES

Abstract.

Background. The research deals with distributed databases of computer-aided design systems with different data structures. The subject of the research is a process of knowledge discovery from these databases. The purpose of the study is to develop a subsystem's structure of knowledge discovery from distributed heterogeneous databases.

Materials and methods. The state of distribution of data sources, the heterogeneity of data within them and the computational complexity of the analysis of large data stipulate implementation of the agent-based approach to achieving this goal.

Results. The organizational model of a multi-agent system of knowledge discovering from distributed heterogeneous databases is developed. The basic roles of agents and their interactions with each other are described.

Conclusions. The main part of the architecture of a subsystem of knowledge discovering from distributed heterogeneous sources is determined by two subsystems: a dataset preparing subsystem and a data mining subsystem. The main problems in

development of this sub-class are due to different structures of data presented in local sources, as well as different accuracy, reliability and completeness of data.

Key words: data base, data mining, CAD, multi-agent systems, knowledge discovery, data fusion.

Введение

Широкое распространение автоматизированных информационных систем в различных сферах человеческой деятельности порождает огромное количество данных [1]. При проектировании сложных объектов в базах данных (БД) систем автоматизированного проектирования (САПР) накапливается большое количество данных, которые после определенного анализа могут быть полезны в будущем при эксплуатации изделия или проектировании усовершенствованных аналогов. Под анализом в данной статье понимается извлечение знаний из данных – итерационный процесс извлечения ранее неизвестных, практически полезных и доступных интерпретаций знаний из наборов данных, ядром которого являются методы интеллектуального анализа данных (ИАД) [1–3]. В качестве знаний могут выступать: кластеры, ассоциативные правила, продукционные правила, математические модели, графы решений, нейронные сети и др.

В условиях современного производства на этапе проектирования сложного изделия может быть задействовано несколько территориально отдаленных друг от друга организаций, использующих специализированные САПР для решения некоторой части общей задачи. Данные в таком случае хранятся в различных распределенных структурированных (базы данных, хранилища данных и др.) и неструктурированных (Интернет, текстовые файлы и др.) источниках. Извлечение знаний из подобных источников является сложной задачей ввиду [4]:

– различной структуры данных в локальных источниках;

– неполноты, противоречивости и других ошибок в данных;

– различной физической природы данных, различной точности и надежности;

– больших объемов и размерности данных;

– сложности извлечения знаний из неструктурированных источников.

К настоящему времени известны три подхода к извлечению знаний из распределенных источников [5]. В одном из них в процесс ИАД вовлекаются все данные из локальных источников, объединенные в единый набор данных. В другом подходе алгоритмы ИАД применяются независимо к локальным наборам данных с последующим объединением частных результатов. Третий подход, который является объединением первых двух, можно разделить на четыре этапа:

1) объединение локальных наборов данных в единый набор данных;

2) поиск и устранение ошибок в наборе данных;

3) распараллеливание вычислений по задачам или данным;

4) объединение частных результатов.

В данной работе рассматривается третий из названных подходов, поскольку он позволяет получить более качественные результаты за приемлемое время за счет распределения вычислений и более объемной обучающей выборки, очищенной от найденных ошибок.

Распределенность источников данных и распределенность вычислений при ИАД обусловливают выбор многоагентного подхода [6–8] к извлечению знаний, в котором каждый отдельный агент имеет частичное представление о задаче и способен решить некоторую ее часть. При этом агенты, взаимодействуя между собой, способны решать сложную задачу.

Согласно методологии Gaia [9] разработка прикладных многоагентных систем (МАС) состоит из двух этапов: этап анализа предметной области и этап проектирования прикладной системы. При этом целью этапа анализа является достижение понимания системы и ее структуры без описания каких-либо деталей разработки, а целью этапа проектирования – трансформация абстрактных решений и понятий, описанных на этапе анализа, в модели более низкого уровня абстракции, которые затрагивают уже описание деталей разработки.

Целью данной статьи является первый этап разработки МАС – анализ предметной области, на котором решается задача построения организационной модели МАС, состоящая из описания моделей ролей агентов и описания их взаимодействия.

1. Анализ предметной области

Процесс извлечения знаний из БД согласно методологии, предложенной Григорием Пятецким – Шапиро и Усама Файадом [10], можно разделить на четыре этапа:

– постановка задачи извлечения знаний;

– подготовка набора данных для анализа;

– ИАД;

– сохранение, применение и визуализация извлеченных знаний.

На этапе постановки задачи определяются цель извлечения знаний, источники данных для извлечения знаний, вид требуемых знаний и требования к их качеству.

На следующем этапе решается задача подготовки качественного набора данных для анализа, которая заключается в сборе данных из различных источников, их объединения и обогащения. Основная проблема на данном этапе обоснована гетерогенностью источников данных, которая заключается в необходимости обеспечения глобальной однозначности семантики терминов, используемых при спецификации данных локальных источников [4].

Следствием того, что набор данных для анализа формируется путем объединения данных из нескольких источников, является наличие в нем большого количества ошибок, которые могут повлиять на производительность алгоритмов анализа и качество извлеченных знаний. Для получения адекватных знаний из имеющихся данных применяются методы по их первичной обработке. Они включают в себя две стадии: обнаружение и устранение ошибок. На первом шаге данные исследуются на предмет «загрязненности», устанавливается, есть ли в них ошибки и к какому виду они относятся. В зависимости от обнаруженных недостатков на следующем шаге применяются различные алгоритмы очистки, объединения дубликатов, устранения противоречий, заполнения пропущенных значений и др.

После первичной обработки набора данных производится выбор задачи ИАД. Выделяют две главные задачи ИАД: прогнозирование (классификация,

регрессия) и описание (кластеризация, визуализация и поиск ассоциативных правил). Для сравнения работы алгоритмов и выбора оптимального результата, в зависимости от задачи анализа, выбирается один или несколько алгоритмов ИАД.

Каждый алгоритм ИАД имеет свои особенности применения, зависящие от имеющихся исходных данных, вычислительных ресурсов, а также от требуемого вида и качества получаемых закономерностей. В целях преобразования набора данных к определенному представлению, формату или виду, оптимальному с точки зрения применяемого алгоритма, производится вторичная обработка данных, к которой относятся процессы:

– поиска аномальных значений;

– кодирования и нормализации;

– выбора наиболее весомых (значимых) атрибутов;

– понижения размерности пространства атрибутов;

– группировки, фильтрации и агрегирования;

– подготовки обучающей и тестовой выборки для решения задачи прогнозирования.

На выходе процесса подготовки набора данных образуется набор данных, приведенный к формату, оптимальному с точки зрения решаемой задачи, очищенный от найденных ошибок и пригодный для применения выбранных алгоритмов ИАД.

На этапе ИАД выбранные алгоритмы могут применяться несколько раз с различными параметрами до тех пор, пока не будет достигнут желаемый результат. Для определения качества примененного алгоритма ИАД на исходном наборе данных производится оценка и интерпретация полученных зависимостей с учетом поставленных на первом этапе целей. На данном этапе выполняется проверка извлеченных знаний на их адекватность и значимость – не являются ли полученные знания случайными для использованного набора данных и нехарактерными для данной предметной области в целом. Для получения независимой оценки адекватности модели выполняется тестирование на новых данных, не участвующих при ее построении, при этом данные, применяемые для тестирования, должны соответствовать сигнатуре модели.

Заключительным этапом является применение и визуализация извлеченных знаний, на котором полученные знания применяются для решения новых задач путем подачи на вход модели некоторых исходных данных и получения приемлемого результата на выходе. При применении модели пользователь должен иметь возможность уточнить, какая информация ему нужна при применении модели. Это позволит оптимизировать применение модели к новым данным.

2. Описание моделей ролей и их взаимодействия

Многоагентная архитектура системы извлечения знаний из распределенных гетерогенных баз данных САПР состоит из: компонент, которые отвечают за отдельные этапы извлечения знаний, рассмотренные в разд. 1, и компонент, отвечающих за координацию работы всей системы (рис. 1).

На первом этапе подготовки набора данных для анализа – постановки цели извлечения знаний – МАС состоит из двух координирующих агентов (рис. 2) – пользователя и интерфейсного агента.

Рис. 1. Организационная структура МАС извлечения знаний из БД

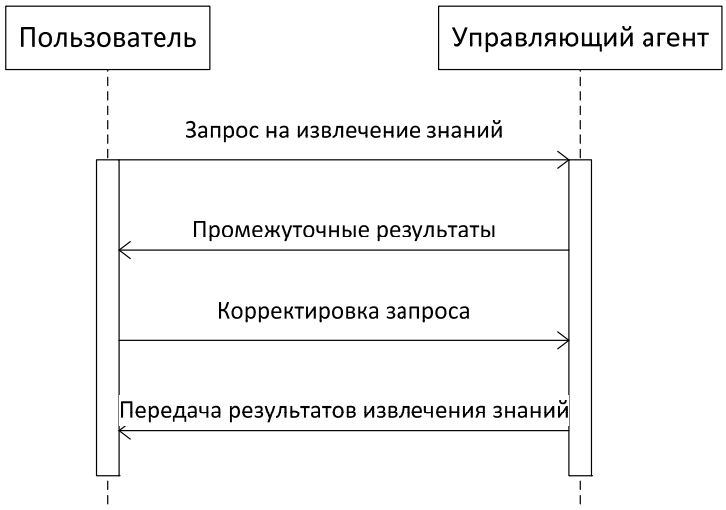

Рис. 2. Взаимодействие пользователя и управляющего агента

Пользователь – эксперт в предметной области – формирует цель извлечения знаний, передает запрос интерфейсному агенту на извлечение знаний и анализирует промежуточные результаты работы отдельных этапов извлечения знаний.

Интерфейсный агент (*управляющий агент*) – интеллектуальный агент, в задачи которого входит получение запросов от пользователя, управление процессом извлечения знаний и предоставление пользователю промежуточных результатов работы МАС.

2.1. МАС сбора, объединения и обогащения данных

На этапе сбора, объединения и обогащения данных МАС состоит из (рис. 3):

– *агента онтологий* – ответственен за анализ и сопоставление структур данных локальных источников с терминами онтологии предметной области;

– *агентов управления БД* – обладает информацией о структуре данных локального источника и исполняет роль шлюза, через который предоставляется доступ к данным локального источника;

– *агентов поиска* – ответственны за последовательный (один агент) или параллельный (несколько агентов) поиск данных в локальных источниках. Достоинством параллельного поиска данных является высокая производительность, так как поиск осуществляется одновременно во всех базах данных несколькими агентами. К недостатку параллельного извлечения данных можно отнести высокую нагрузку на сеть в момент передачи набора данных от поискового агента управляющему агенту;

– *агента объединения данных* – ответственен за объединение нескольких наборов данных в один.

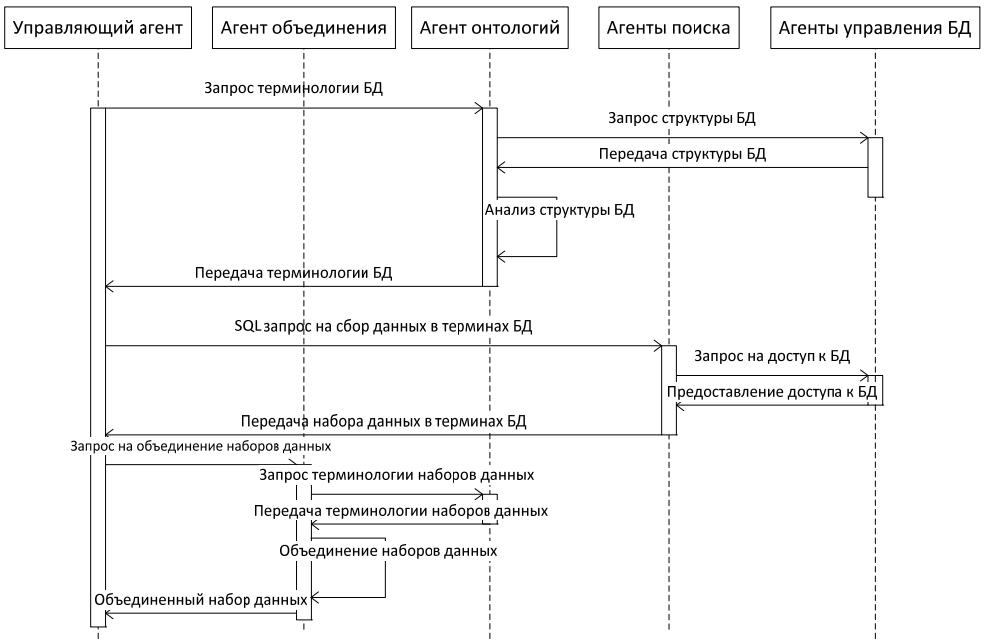

Рис. 3. МАС сбора и объединения данных для анализа

Задача внешнего обогащения набора данных может решаться МАС сбора и объединения данных с отличными от первоначальных критериями поиска и/или источниками данных.

2.2. МАС первичной обработки набора данных

На этапе первичной обработки набора данных МАС (рис. 4) состоит из:

– *агентов поиска ошибок* – ответственны за последовательный (один агент) или параллельный (несколько агентов) поиск выбранных ошибок в наборе данных;

– *агентов устранения ошибок* – ответственны за последовательное (один агент) или параллельное (несколько агентов) устранение выбранных ошибок;

– *агента объединения результатов* – ответственен (при параллельном устранении ошибок) за объединение частных результатов устранения ошибок в единый набор данных.

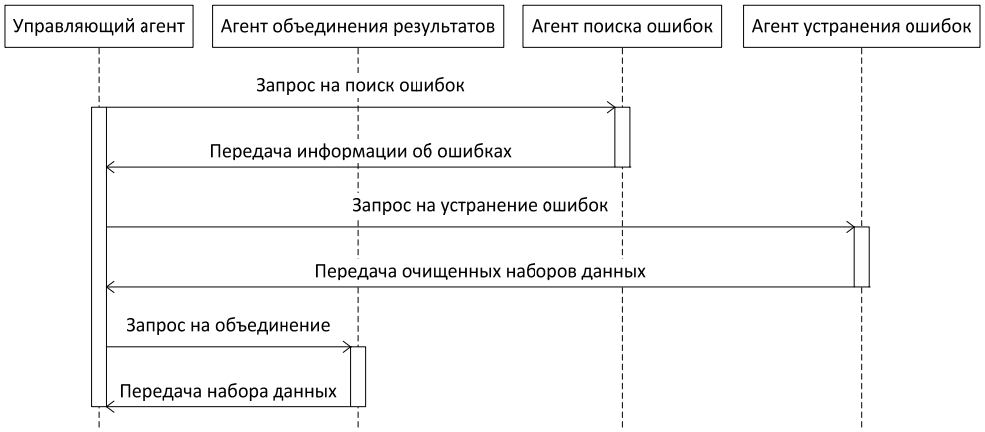

Рис. 4. МАС первичной обработки набора данных

Стоит отметить, что на выходе качество набора данных может отличаться в зависимости от последовательности, в которой производится устранение каждого из видов ошибки. Для повышения качества набора данных следует провести первичную обработку с различной последовательностью устранения ошибок с целью определения оптимальной последовательности для конкретного набора данных. Критерием оценки качества в данном случае может выступать экспертная оценка качества наборов данных.

2.3. МАС вторичной обработки набора данных

На этапе вторичной обработки набора данных (рис. 5) управляющий агент передает агенту вторичной обработки запрос на обработку набора данных. Количество агентов вторичной обработки определяется количеством алгоритмов, выбранных пользователем для анализа.

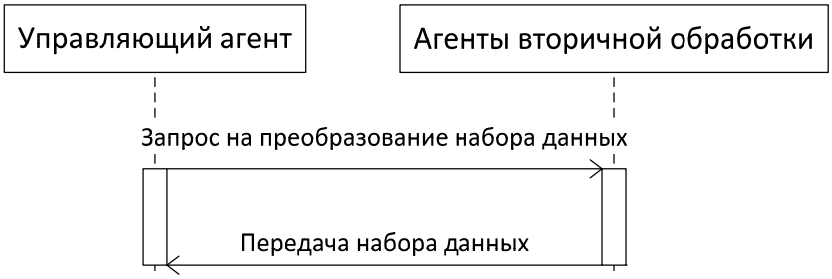

Рис. 5. МАС вторичной обработки набора данных

Агенты вторичной обработки данных получают информацию от управляющего агента о том, в каком виде должен быть представлен набор данных,

производят с ним соответствующие изменения и передают обратно для дальнейшего анализа.

2.4. МАС интеллектуального анализа данных

На этапе интеллектуального анализа данных (рис. 6) МАС состоит из:

– *агентов ИАД* – ответственны за последовательный (один агент) или параллельный (несколько агентов) анализ набора данных (в соответствии с параметрами одного или нескольких выбранных алгоритмов), декомпозицию вычислений и объединение результатов распределенных вычислений;

– *вычислительных агентов* – предоставляют агенту ИАД свои вычислительные ресурсы;

– *агента сравнения результатов анализа* – ответственен за сравнение результатов анализа по выбранному критерию;

– *агента управления базой знаний* – предоставляет доступ управляющему агенту к базе знаний для сохранения извлеченных знаний.

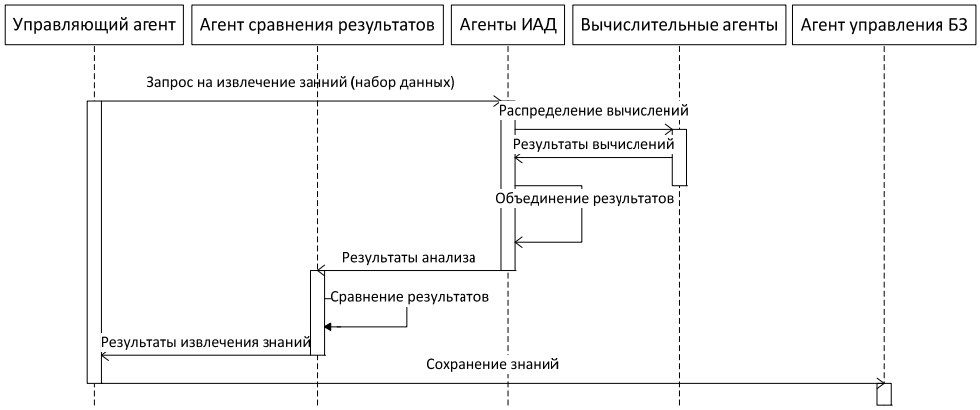

Рис. 6. МАС ИАД

Заключение

По существу основная часть архитектуры подсистемы извлечения знаний из распределенных источников определяется подсистемой подготовки набора данных и подсистемой ИАД. Основные проблемы при проектировании подсистем такого класса обусловлены распределенностью источников данных, гетерогенностью данных, представленных в источниках, и вычислительной сложностью анализа больших объемов данных. Для разработки системы, решающей проблемы подобного класса, в данной работе предлагается использовать агентно-ориентированный подход. В работе произведен первый этап согласно методологии Gaia разработки многоагентной системы – анализ предметной области, который заключается в построении организационной модели, определении ролей агентов и взаимодействия между ними.

Конечной целью данной работы является автоматизированное наполнение базы знаний интеллектуальной САПР. Извлеченные из базы данных САПР знания могут быть использованы для принятия проектных решений при разработке новых изделий методом совершенствования аналогов и заимствования удачных проектных решений.

Список литературы

1. **Maimon, O.** Data Mining and Knowladge Discovery Handbook / O. Maimon, L. Rokach. – 2nd ed. // Springer, Science+Business Media, 2010. – 1285 p.
2. **Han, J.** Data Mining: Consepts and techniques / J. Han, M. Kamber. – 2nd ed. – Morgan Kaufmann, 2006. – 743 p.
3. **Larose, D. T.** Data mining methods and models / Daniel T. Larose. – Wiley-IEEE Press, 2006. – 344 p.
4. **Городецкий, В.** Многоагентная технология принятия решений в задачах объединения данных / В. Городецкий, О. Карсаев, В. Самойлов // Труды СПИИРАН. – 2003. – № 1. – С. 12–37.
5. **Куприянов, М. С.** Интеллектуальный анализ распределенных данных на базе облачных вычислений / М. С. Куприянов. – СПб. : Изд-во СПБГЭТУ «ЛЭТИ», 2011. – 148 с.
6. **Шевцов, А. Н.** Агентно-ориентированные системы: основные модели : моногр. / А. Н. Шевцов. – Вологда : ВоГТУ, 2012. – 189 с.
7. **Ющенко, С. П.** Многоагентные системы информационной поддержки управленческих решений / С. П. Ющенко. – Ростов н/Д. : Изд-во СКНЦ ВШ, 2004. – 376 p.
8. **Тарасов, В. Б.** От многоагентных систем к интеллектуальным организациям: философия, психология, информатика / В. Б. Тарасов. – М. : Эдиториал УРСС, 2002. – 352 с.
9. **Wooldridge, M.** The Gaia Methodology for Agent-Oriented Analysis and Design / M. Wooldridge, N. R. Jennings, D. Kinny // Journal of Autonomous Agents and Multi-Agent Systems. – 2000. – Vol. 3, № 3. – P. 285–312.
10. **Usama Fayyad.** From Data Mining to Knowledge Discovery in Databases / Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth // AI Magazine. – 1996. – Vol. 17, № 3. – P. 37–54.

References

1. Maimon O., Rokach L. *Data Mining and Knowladge Discovery Handbook. 2nd ed.* Springer, Science+Business Media, 2010, 1285 p.
2. Han J., Kamber M. *Data Mining: Consepts and techniques*. 2nd ed. Morgan Kaufmann, 2006, 743 p.
3. Larose D. T. *Data mining methods and models*. Wiley-IEEE Press, 2006, 344 p.
4. Gorodetskiy V., Karsaev O., Samoylov V. *Mnogoagentnaya tekhnologiya prinyatiya resheniy v zadachakh ob"edineniya dannykh* [Multi-agent technology of decision making in pooled data problems]. SPIIRAS Proceedings. 2003, no. 1, pp. 12–37.
5. Kupriyanov M. S. *Intellektual'nyy analiz raspredelennykh dannykh na baze oblachnykh vychisleniy* [Intelligent analysis of distributed data on the basis of cloud computing]. Saint-Petersburg: Izd-vo SPBGETU «LETI», 2011, 148 p.
6. Shevtsov A. N. *Agentno-orientirovannye sistemy: osnovnye modeli: monogr.* [Agent-oriented systems: basic models: monograph]. Vologda: VoGTU, 2012, 189 p.
7. Yushchenko S. P. *Mnogoagentnye sistemy informatsionnoy podderzhki upravlencheskikh resheniy* [Multi-agent systems of managerial decision information support]. Rostov-on-Don: Izd-vo SKNTs VSh, 2004, 376 p.
8. Tarasov V. B. *Ot mnogoagentnykh sistem k intellektual'nym organizatsiyam: filosofiya, psikhologiya, informatika* [From multi-agent systems towards intelligent organizations: philosophy, psychology, informatics]. Moscow: Editorial URSS, 2002, 352 p.
9. Wooldridge M., Jennings N. R., Kinny D. *Journal of Autonomous Agents and Multi-Agent Systems*. 2000, vol. 3, no. 3, pp. 285–312.
10. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth *AI Magazine*. 1996, vol. 17, no. 3, pp. 37–54.

Бондаренко Игорь Борисович
кандидат технических наук, доцент, кафедра проектирования и безопасности компьютерных систем, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Россия, г. Санкт-Петербург, Кронверкский пр. 49)

E-mail: igorlitmo@rambler.ru

Bondarenko Igor Borisovich
Candidate of engineering sciences, associate professor, sub-department of computer system engineering and safety, Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics (49 Kronverksky avenue, Saint-Petersburg, Russia)

Иванов Алексей Игоревич
аспирант, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Россия, г. Санкт-Петербург, Кронверкский пр. 49)

E-mail: 145732@niuitmo.ru

Ivanov Aleksey Igorevich
Postgraduate student, Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics (49 Kronverksky avenue, Saint-Petersburg, Russia)

УДК 004.896
Бондаренко, И. Б.

Организационная модель многоагентной системы извлечения знаний из распределенных гетерогенных баз данных систем автоматизированного проектирования / И. Б. Бондаренко, А. И. Иванов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2015. – № 4 (36). – С. 54–63.