УДК 004.415.5:004.896

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ПРОИЗВОДСТВЕННЫХ ДАННЫХ

А.В. Кудинов

Томский политехнический университет
E-mail: KudinovAV@tpu.ru

Предложена оригинальная технология решения задач интеллектуального анализа производственных данных промышленных предприятий. Описана концепция создания такой технологии, включая типовую архитектуру BI-системы, подходы к проектированию и реализации центрального хранилища производственных данных, тематических витрин данных, а также способ реализации моделей анализа производственных данных на основе технологии Data Mining. Приведен пример решения прикладной задачи анализа на реальных данных промышленного предприятия.

Ключевые слова:
Аналитическая обработка данных, интеллектуальный анализ данных, хранилища данных, информационно-аналитические системы.

Key words:
On-line analytical processing, data mining, data warehouses, business intelligence.

Введение

При автоматизации любого производственного бизнес-процесса наступает этап, когда проблемы сбора данных, первичные учетные и оперативные задачи уже, в целом, решены, накоплен достаточный массив информации и на первый план выходят проблемы анализа этих данных. При этом анализ необходим как для решения повседневных задач (например, прогнозов необходимости ремонта оборудования по его состоянию или необходимости закупки тех или иных запчастей или материалов в зависимости от степени их текущего использования), так и для определения стратегического направления развития компании в целом.

Таким образом, актуальной становится задача выбора подходящих средств для анализа и поддержки процесса принятия решений. В этом случае традиционно применяются такие инструменты, как *хранилища данных* (Data Warehouses, ХД) и построенные на их основе *системы бизнес-анализа данных* (Business Intelligence, BI), в российской научной литературе более известные как *информационно-аналитические системы* или *системы поддержки принятия решений*. Фундаментом средств бизнес-аналитики являются технологии OLAP и Data Mining [1, 2].

Традиционными сферами применения вышеперечисленных технологий является анализ экономической (финансовой) информации. В то же время, растет интерес к BI, технологиям OLAP и Data Mining в социальной сфере (образование, медицина, государственное управление и т. д.), а также в промышленности (особенно в крупной – энергетика, металлургия, нефтехимия, нефтегазодобыча) [2–5]. Причем в сфере их внимания оказывается не только решение задач финансово-экономического анализа, так или иначе актуальных для всех предприятий и организаций, но и к решение сугубо специфичных, отраслевых аналитических задач.

При этом недостаточно развитой остается теоретическая база, связанная с научным обоснованием создания систем анализа производственных данных. Большинство исследователей сосредоточено либо на решении конкретных прикладных задач отдельных отраслей и компаний, либо, наоборот, на определении возможности использования конкретных групп аналитических методов для широкого круга производственных задач. Из работ зарубежных исследователей можно выделить монографии [3, 4], представляющие наиболее полные аналитические обзоры методов прикладной математики для решения задач управления производственными процессами различных отраслей. Из работ российских исследователей можно отметить [6–8], посвященные развитию методов и созданию прикладных систем для решения аналитических задач для нефтеперерабатывающих производств.

К сожалению, часто за границами внимания исследователей остаются важные проблемы применения интеллектуального анализа в промышленности, связанные с недостаточной квалификацией пользователей и сложностью разработки типовых решений, универсальных для целых групп отраслей. По мнению авторов, снять остроту вышеперечисленных проблем и снизить издержки при разработке и внедрении систем бизнес-анализа в производстве может создание комплексной *информационной технологии для решения задач интеллектуального анализа производственных данных.*

Концепция создания технологии решения задач интеллектуального анализа производственных данных

Можно выделить следующие аспекты, усложняющие разработку и внедрение систем интеллектуального анализа в производственной деятельности промышленных предприятий:

1) слабая готовность пользователей к применению сложных математических методов;
2) высокая стоимость и длительные сроки внедрения BI-систем;

3) сложность масштабирования таких систем либо для решения новых задач, либо при их адаптации к изменяющейся производственной ситуации.

Рассмотрим эти аспекты более подробно.

Наиболее серьезной проблемой внедрения BI-систем в производстве, на наш взгляд, является сложность для конечных пользователей как математического аппарата и программных средств анализа, так и инструментов для визуализации их результатов. Большинство известных аналитических методов и популярных инструментов оперируют концепциями, далекими от задач предметной области. Даже квалифицированный геолог или технолог не сможет сразу найти соответствие между актуальной для него производственной задачей и определенным классом задач анализа (кластеризацией, ассоциацией, анализом отклонений или последовательностей), не говоря о том, чтобы выбрать подходящий метод (алгоритм) решения, подобрать его параметры и затем правильно интерпретировать качество решения задачи, основываясь на предлагаемых статистических характеристиках результатов.

С другой стороны, специалисты по анализу данных, имеющие необходимую квалификацию для его проведения, не всегда являются специалистами в предметной области. Таким образом, они не способны как к выявлению места аналитических задач в производственных бизнес-процессах, так и обоснованному оперированию параметрами методов и, в конечном итоге, к правильной интерпретации результатов с точки зрения их полезности для бизнеса. В этой связи «идеальным» пользователем такой системы будет высококвалифицированный специалист, сочетающий в себе знания методов и инструментов анализа и опыт решения прикладных задач в предметной области. Именно такие специалисты наиболее востребованы в традиционных для BI сферах применения, таких как экономический, финансовый анализ и др. Многие компании создают даже специальные аналитические службы, работающие в интересах бизнеса в целом. Однако в производственной сфере найти или подготовить аналитиков, способных также работать на стыке нескольких технических специальностей, гораздо сложнее.

Нам представляется более продуктивным другой подход, согласно которому для полноценного использования методов и технологий интеллектуального анализа в производственной сфере необходимо найти способ приблизить данные технологии к потребностям конкретных специалистов. Данная проблема усугубляется тем, что разработка информационно-аналитических систем и проектирование архитектуры хранилища данных сами по себе являются сложными, дорогостоящими и длительными процессами. Типичный проект внедрения BI-системы (по данным исследователей) имеет продолжительность от 0,5 до 2 лет и стоимость от 0,4 до 1,5 млн долларов [9]. Поэтому очевидно востребованными на рынке являются типовые решения в этой области, которые аккумулируют опыт разработки и внедрения систем такого рода на десятках и даже сотнях предприятий. К сожалению, сравнительно легкой типизации поддается, в основном, решение аналитических задач, общих для большинства компаний и относящихся, опять же, к традиционным для BI сферам применения: финансовый анализ, анализ рынков, управление персоналом и т. д. Примеров создания и внедрения BI-систем, решающих задачи анализа производственных процессов, настолько мало, что говорить о типовых отраслевых решениях пока рано, хотя актуальность их разработки не вызывает сомнений.

На первом этапе создания технологии решения задач интеллектуального анализа производственных данных необходимо очертить круг собственно аналитических задач, актуальных при управлении производством.

Задачи анализа производственных данных

Сформировать общее видение концепции BI-системы компании и тем более применять разработанные BI-средства невозможно без четкого видения аналитических задач, необходимых как для предприятия в целом, так и для отдельных его подразделений. В работе [8] приведен следующий перечень классов задач для анализа производственных (технологических) ситуаций, связанных с обнаружением и прогнозированием скрытых тенденций и закономерностей развития производственных процессов, выявлением скрытых факторов влияния и взаимосвязей между производственными параметрами, а также оптимизацией управления производственными процессами и визуализацией результатов анализа. В работе [2] этот перечень дополняется такой задачей, как прогнозирование качества изделия в зависимости от некоторых параметров технологического процесса. Также там отмечается, что природа технических процессов достаточно стабильна и их отклонения должны лежать в некоторых определенных пределах, что облегчает проведение некоторых видов анализа этих процессов.

При создании BI-систем для определенных отраслей и производств важно конкретизировать и группы аналитических задач. Обычно они соответствуют задачам определенных групп специалистов (диспетчеров, технологов, механиков и т. д.). Практически идентичной является существенная часть функций, методов работы и используемых данных для целого ряда производственных служб промышленных предприятий различных отраслей. Такая унификация объясняется, в частности, тем, что задачи, методы работы, формы используемых документов каждой службы определяются, в значительной степени, не отраслевой спецификой, а нормами государственного технического контроля и другими общепринятыми нормами и регламентами. В этой связи, нет существенных отличий

между работой службы главного механика, например, металлургического, нефтехимического производства и газодобывающих компаний (ГДК). То же самое относится и к функциям служб главного метролога, энергетика, связиста и т. д. Соответственно в значительной степени инвариантны к отраслевой специфике и аналитические задачи, решаемые этими службами.

Например, газовая отрасль специфична как по своей области знаний, так и по особенностям анализируемых данных, поэтому для определения целей анализа и постановки задач необходим четко проработанный анализ предметной области. Таким образом, можно сформулировать общие и частные аналитические задачи, актуальные при управлении производством ГДК. К общим производственно-аналитическим задачам, решаемым на основе данных всего предприятия, мы предлагаем отнести анализ:

- эффективности выполнения основных и вспомогательных бизнес-процессов;
- режимов работы основного и вспомогательных производств (разведка и бурение, добыча, подготовка/переработка, транспортировка), в том числе выявление аномальных режимов, прогноз и оптимизация;
- выполнения производственных планов и достижения основных производственных показателей, в том числе на основе систем KPI, BSC и т. д.;
- состояния производственного оборудования, в том числе для оптимизации управления его ремонтами;
- планов развития производственной базы;
- эффективности использования ресурсов в производстве, в том числе анализ эффективности системы материально-технического снабжения;
- качества сырья и продукции;
- квалификации и эффективности работы производственного персонала.

Разработка типовой архитектуры BI-системы

Создание технологии решения аналитических задач подразумевает разработку типовой архитектуры BI-системы для анализа производства, ее математического и программного обеспечения, а также методологий ее внедрения и сопровождения.

В основе архитектуры предлагается использовать классический подход к построению BI-системы, предполагающий создание как центрального хранилища данных, так и структур более мелких *витрин данных*, предназначенных для решения аналитических задач отдельных подразделений или групп специалистов предприятия. При этом данные центрального хранилища должны накапливаться непрерывно в процессе производства в течение долгого периода времени. Они представляют собой квинтэссенцию опыта специалистов предприятия, сосредоточение знаний о характере протекания производственных процессов, что делает их неоценимой основой для поддержки принятия управленческих решений, оптимизации производственных процессов в масштабах всего предприятия и пр. Для решения частных аналитических задач отдельных групп специалистов огромный массив данных всего предприятия требуется декомпозировать на подмножества данных витрин. На рис. 1 приведена предложенная общая архитектура BI-системы для анализа производства. Ее главным элементом является центральное хранилище производственных данных (ЦХПД), источниками данных которого являются различные транзакционные системы, так или иначе оперирующие производственно-технологическими данными. Поступление данных из источников управляется инте-

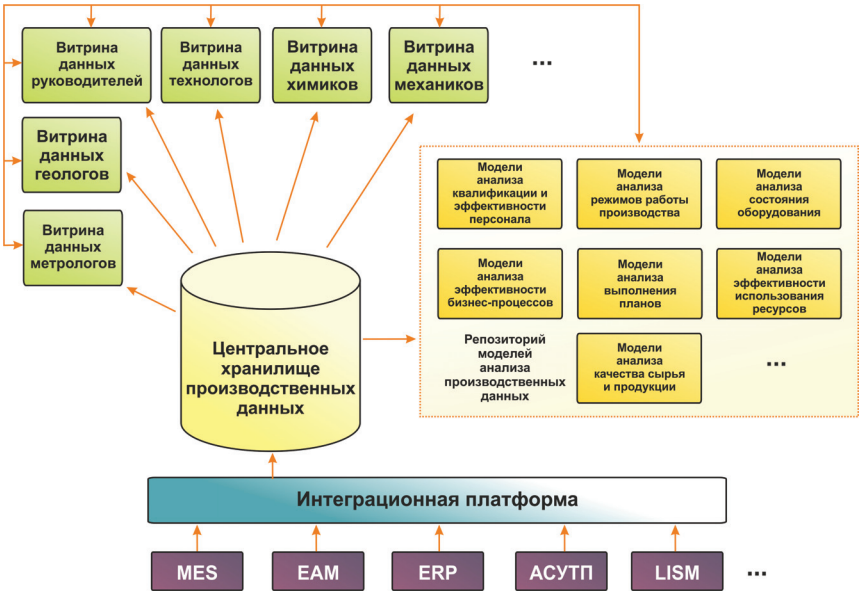

Рис. 1. *Общая архитектура BI-системы для анализа производства*

грационной платформой, подробно описанной в работе [10].

Функции системы доступны ее пользователям через тематические витрины данных, ряд задач которых решается на основе моделей из общего репозитория моделей анализа производственных данных. Основным принципом предложенной архитектуры является легкая масштабируемость. Изменение состава и структуры источников данных, появление новых аналитических задач не должно оказывать существенного влияния на структуру центрального хранилища данных и существующих тематических витрин данных. Модели данных как центрального хранилища, так и витрин, должны быть максимально инвариантны к особенностям конкретных предприятий, источников данных и программных платформ.

С точки зрения практической реализации предложенной архитектуры, можно рекомендовать использование любой современной BI-платформы, такой как Oracle BI Suite, Microsoft SQL Server Analyses Services, SAP BusinessObjects BI Platform и др. В данной работе использовалась платформа Microsoft SQL Server 2008 Analyses Services [1], поэтому описание приведенных примеров выполнено с использованием некоторых терминов, концепций и элементов интерфейса именно этого продукта.

При проектировании структуры центрального хранилища производственных данных (ЦХПД) преследуются следующие основные цели:

1) организовать легкое поступление данных в хранилище из учетных систем;
2) организовать легкую выборку данных из хранилища для тематических витрин;
3) по возможности абстрагироваться в структуре хранилища от отраслевых особенностей;
4) организовать анализ данных непосредственно в хранилище, безотносительно функциональных и даталогических рамок, определяемых интересами конкретных групп пользователей (служб) и задаваемых структурами данных и функциями соответствующих витрин;
5) обеспечить возможность легкого масштабирования хранилища при появлении новых учетных систем, витрин и аналитических задач.

Поскольку существуют принципиальные различия в проектировании структур хранилищ данных и транзакционных (учетных) систем [9], то первая и вторая цель фактически противоречат друг другу. Поэтому можно предложить использовать подход, применяемый при создании *оперативных складов данных*, служащих технологическим буфером между учетными системами и хранилищами данных. В соответствии с этим подходом, создают гибридную структуру, позволяющую, с одной стороны, учесть максимум фактических данных учетных систем с той же (по возможности) детальностью и атрибутивной полнотой, а с другой стороны, определить для этих данных модели агрегации, эквивалентные соответствующим моделям хранилищ (витрин) данных. При этом мы предлагаем группировать фактические данные в хранилище по типам производственных данных, т. е. выделить плановую, паспортную, оперативно-технологическую, пространственную и иную информацию в отдельные кубы данных. Таким образом, достигаются и цели абстрагирования от особенностей предметной области, и относительно легкого масштабирования структуры ЦХПД. Кроме того, группировка по типам производственной информации позволяет проводить на этих данных недетерминированный, разведочный их анализ, что также является одной из обозначенных нами целей. Необходимо отметить, что такой принцип построения ЦХПД не позволяет применять его конечными пользователями в качестве легкого инструмента для решения повседневных прикладных задач анализа, т. к. его структура в большей степени определяется особенностями управления данными, чем задачами и объектами конкретной предметной области.

Витрина данных и названа так потому, что разворачивает информацию и функции BI-системы лицом к конечному пользователю. Поэтому основными требованиями к ее функциям и структуре ее данных являются простота понимания и использования. Для именования элементов структуры данных витрины необходимо использовать термины предметной области. То же требование можно отнести, по возможности, и к структурам и моделям анализа данных (*Data Mining Structures* и *Data Mining Models*), используемым витриной (в предлагаемой нами архитектуре они вынесены в отдельный репозиторий моделей анализа). Так, для витрины геолого-промысловых данных фактами будут результаты исследований скважин, объемы извлечения сырья, данные о ходе бурения, факты остановок скважин и проведения геолого-технических мероприятий. Измерениями, соответственно, будут объекты разработки (месторождения, пласты, блоки, скважины, кусты скважин) и время.

Разработка моделей анализа производственных данных и примеры решения прикладных задач анализа

В качестве методологической основы для разработки моделей анализа производственных данных мы предлагаем использовать широко известную методологию CRISP-DM [11]. В качестве отличительной черты предлагаемой в данной работе технологии можно отметить широкое использование шаблонов, как инструмента для масштабирования BI-системы в условиях возникновения новых аналитических задач, а также репозитория моделей анализа, как важного элемента архитектуры системы. Соответственно, шаблоны позволяют легко решать новые аналитические задачи на основе уже реализованных задач аналогичного класса (но, например, отличающихся областью данных), а общий репозиторий позволяет использовать эти шаблоны различными производственными службами, не ограничиваясь рамками тематических витрин.

В качестве примера решения прикладной задачи анализа производственных данных при помощи предложенной технологии приведем результаты исследования по выявлению ассоциативных связей между возникновением недостоверных данных измерительных устройств и их характеристиками. Моделирование выполнялось на основе реальных данных архива технологических параметров MES «Магистраль-Восток», внедренной в ОАО «Востокгазпром» [12]. Для технологического процесса нефтегазодобычи характерно наличие возмущений и искажений в значениях технических параметров, контролируемых при помощи автоматизируемых систем управления технологическими процессами (АСУ ТП). Это обусловлено динамикой электрических процессов, происходящих в агрегатах на объектах нефтедобычи. Случайные паразитные явления, такие как механические вибрации, образование газовых скоплений, неоднородность закачиваемого вещества в реальной ситуации оказываются причиной появления пиков и пульсаций в замерах технологических параметров. При сборе технологических данных от АСУ ТП по протоколу OPC (OLE for Process Control) соответствующий OPC-сервер может возвращать вместе со значением самого технологического параметра также значение атрибута *Достоверность техпараметра* (как одного из состояний – *Good*, *Bad* или *Uncertain*). Для ГДК общее число технологических параметров может составлять до нескольких десятков тысяч, причем исторический архив может содержать до нескольких миллионов значений каждого параметра. Для отдельных параметров техпараметра возможно посчитать количество значений с *Bad* достоверностью, соотнести это количество к общему количеству всех обработанных значений (*Bad, Good, Uncertain*), т. е. создать оценку доли недостоверных значений в общем количестве значений параметра. Таким образом, задачей является выявление сочетания таких классификационных признаков, которые помогают определить проблемные места, используя BI-средства.

В результате в качестве классификационных признаков технологических параметров были выбраны: название АСУ ТП, тип параметра (давление, температура, расход и т. д.), тип объекта-источника параметра (насосы, задвижки, емкости и т. д.). Была построена модель анализа, использующая упрощенный алгоритм Байеса и алгоритм дерева решения. Результаты моделирования при помощи обоих алгоритмов совпали и показывают, что порядка 80 % всех недостоверных значений дают параметры, контролируемые датчиками температуры АСУ ТП «RS3», установленными на техпозициях насосов Мыльджинской УКПГ. После сужения области поиска было решено провести расширенное исследование и выявить корреляцию между количеством отказов данного типа датчиков и их паспортными характеристиками (дата выпуска, завод-изготовитель, дата ввода в эксплуатацию и т. д. – всего несколько десятков атрибутов).

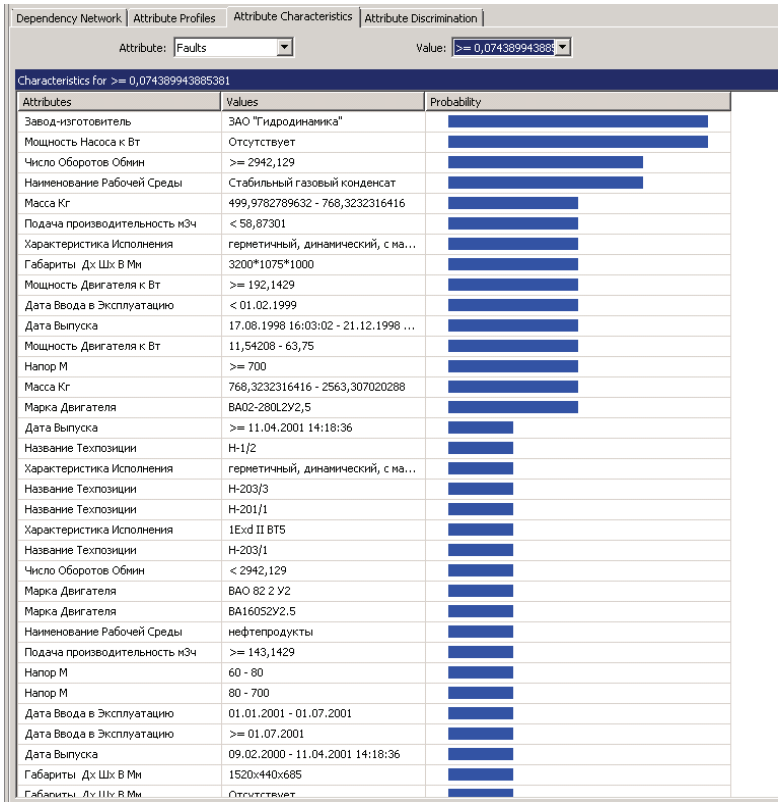

Рис. 2. *Результаты анализа характеристик средств измерения*

При моделировании также использовался упрощенный алгоритм Байеса. Результатами стало выделение такого набора классификационных признаков, который указывал на наибольшую вероятность появления недостоверных значений параметров. В частности выяснилось, что наиболее вероятен отказ датчиков температуры, выпущенных ЗАО «Гидродинамика» в определенный период (рис. 2). Таким образом, можно говорить о статистически подтверждённом факте обнаружения бракованной партии средств измерений, что служит основой для принятия решений специалистами службы главного метролога предприятия. Оценка точности моделирования, выполненная также при помощи средств Microsoft SQL Server 2008 Analyses Services, показала высокую достоверность его результатов. В дальнейшем планируется на основе созданной модели анализа реализовать постоянно действующую модель, пересчет которой бы осуществлялся автоматически при поступлении новых данных с систем нижнего уровня, а нотификация пользователей о найденных результатах выполнялась бы посредством рассылки отчета по электронной почте.

Выводы

Рассмотрена информационная технология для решения широко спектра задач интеллектуального анализа производственных данных промышленного предприятия. Ее основным преимуществом является ориентированность на создание типовых решений для целых отраслей и групп отраслей промышленности. В частности показано, как при проектировании архитектуры BI-системы унифицировать структуру центрального хранилища производственных данных, а также как проектировать структуры витрин данных для конкретных групп специалистов. Предложен подход к решению аналитических производственных задач на основе шаблонов, а также показаны преимущества введения такого элемента архитектуры BI-системы, как единого репозитория моделей анализа.

Рассмотрены результаты применения данной технологии при разработке и внедрении прикладных систем анализа для предприятий газовой отрасли и при решении конкретных аналитических задач производственных служб газодобывающей компании. Так, проведенное исследование по выявлению ассоциативных связей между возникновением недостоверных данных измерительных устройств и их характеристиками показало способность созданных моделей к выявлению партий бракованных датчиков. Проведенное исследование является классическим примером применения технологии Data Mining, т. к. осуществлялась не статистическая проверка заранее сформулированной гипотезы, а поиск нового знания о предметной области в огромном массиве первичных данных.

Работа выполнялась по тематике госбюджетной НИР в рамках государственного задания «Наука».

СПИСОК ЛИТЕРАТУРЫ

1. Макленнен Дж., Танг Чж., Криват Б. Microsoft SQL Server 2008: Data Mining – интеллектуальный анализ данных / Пер. с англ. А. Лашкевича. – СПб.: БХВ-Петербург, 2009. – 720 с.: ил.
2. Чубукова И.А. Data Mining. 2-е изд., испр. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2010. – 382 с.: ил.
3. Feldman R., Curry G.G. Manufacturing Systems Modeling and Analysis. – Berlin, Heidelberg: Springer-Verlag, 2011. – 335 p.
4. Wang Jun. Computational Intelligence in Manufacturing Handbook. – Boca Raton: CRC Press LLC, 2001. – 560 p.
5. Попова М. Во что вложится нефтегаз? Мнения ИТ-экспертов. URL: http://www.3dtvv.ru/index.php? option=com_content&view=article&id=73&lang=ru (дата обращения: 20.01.2012).
6. Гершберг А.Ф., Мусаев А.А., Нозик А.А., Шерстюк Ю.М. Концептуальные основы информационной интеграции АСУ ТП нефтеперерабатывающего предприятия. – СПб.: Альянс-строй, 2003. – 128 с.
7. Гершберг А.Ф. Интеграция и интеллектуализация АСУ ТП нефтеперерабатывающего предприятия. – СПб: ПО «Киришинефтеоргсинтез», 2001. – 52 с.
8. Мусаев А.А. Алгоритмы аналитического управления производственными процессам // Автоматизация в промышленности. – 2004. – № 1. – С. 30–35.
9. Туманов В.Е. Проектирование хранилищ данных для систем бизнес-аналитики. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2010. – 615 с.: ил.
10. Veyber V., Kudinov A., Markov N. Model-driven platform for oil and gas enterprise data integration // International Journal of Computer Applications. – 2012. – № 49 (5). – P. 14–19.
11. Azevedo A., Santos M.P. KDD, SEMMA and CRISP-DM: A parallel overview // IADIS European Conference Data Mining. – Amsterdam, 2008 July 24–28. – P. 182–185.
12. Богдан С.А., Кудинов А.В., Марков Н.Г. Опыт внедрения MES «Магистраль-Восток» в нефтегазодобывающей компании // Автоматизация в промышленности. – 2010. – № 8. – С. 53–58.

Поступила 19.09.2012 г.