Vasyl Kazak, doctor of Technical Sciences, professor, Automation and Energy Management Department, National Aviation University, Komarova Ave, 1, Kiev, Ukraine, 03058
E-mail: profkazak@ukr.net
Dmitro Shevchuk, PhD, associate Professor, Automation and Energy Management Department, National Aviation University. Komarova Ave, 1, Kiev, Ukraine, 03058
E-mail: dmitroshevchuk@gmail.com
Andrii Babenko, PhD, assistant professor, Air Transportation Management Department
National Aviation University, Komarova Ave, 1, Kiev, Ukraine, 03058,
E-mail: andrii.babenko@gmail.com
Mykhailo Levchenko, student, Department of Automation and Energy Management, National Aviation University, pr., Komarova 1, Kyiv, Ukraine, 03058,
E-mail: mhlevchenko@gmail.com

УДК 681.5: 004.5
DOI: 10.15587/2313-8416.2014.29738

ПРИМЕНЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ СИНТЕЗЕ МЕТОДОВ ОЧИСТКИ ДАННЫХ

© В. А. Доровской, С. Г. Черный, И. А. Доровская, Н. П. Сметюх

Осуществлен анализ методов очистки данных мониторинга условий труда; изложены общие принципы и алгоритмы методов очистки данных. Для технологии по очистке данных использование пакетов MICROSOFT SQL SERVER и MATLAB дает положительные результаты, но имеет определенные преграды. В связи с этим возникает необходимость разработки компьютерных промышленных открытых пакетов по методам очистки данных.
Ключевые слова: очистка данных, идентификация, кластеризация, data mining, консолидация, SQL, горно-металлургический, эксперт.

The analysis of methods for monitoring data cleansing of working conditions is proved; the general principles and methods of data cleaning algorithms are given. Usage of MICROSOFT SQL SERVER and MATLAB gives positive results for data cleansing technology, but it has some obstacles. In this regard, there is a need to develop the industrial computer open packages of data cleansing methods.
Keywords: data cleansing, identification, clustering, data mining, consolidation, SQL, ore mining and smelting, expert.

1. Введение

Горно-металлургические предприятия гиганты (горно-обогатительные комбинаты, металлургический комбинат, подземный горнодобывающий комбинат) (ГМП)), которые характеризуются большими объемами производства с использованием значительных человеческих ресурсов, требуют новых ИТ-технологий управления методами очистки данных мониторинга условий труда горно-металлургических рабочих (ГМР) [1, 2]. Мониторинг условий труда ГМР позволяет осуществлять накопление данных оценки УТ с целью дальнейшего использования этих данных при формировании состава пенсионного списка ГМР.

В работе исследуется проблема ИТ-технологий управления методами очистки данных мониторинга УТ ГМР. Актуальность этой проблема очевидна потому что, количество ГМР работающих в очень тяжелых УТ ежегодно увеличивается (табл. 1), что приводит к истощению человеческих ресурсов региона.

2. Анализ литературных источников по очистке данных

Горно-металлургические предприятия относят-ся к сложным динамическим системам [1] и имеют следующие отличительные особенности данных УТ РМ:

– непостоянство РМ;

– необходимость строгого соблюдения планов по выполняемым в определенное время объемам;

– большая зависимость результатов функционирования от природных условий (геологических и климатических);

– трудности в организации функционирования системы постоянно перемещающихся РМ.

С позиции системных компонентов можно выделить: наличие большого количества случайных факторов УТ РМ и трудности прогнозирования протекания технологических процессов, и функционирования всей системы на длительный промежуток времени. С точки зрения аналитического составляющего, заметим, что возникает существенная проблематика в осуществлении автоматизированного управления производственными процессами по причинам непостоянства РМ и отсутствие, необходимых для контроля датчиков, средств передачи информации на управляющий центр, а так же неравномерный режим работы оборудования на протяжении смены и количество данных (табл. 1).

Таблица 1

Характеристика количества трудящихся и их профессий горно-металлургических предприятий Криворожского бассейна

Предприятие	Число трудящихся, профессии
СевГок	16000
ЦГок	12000
НКГок	5000
ЮГок	11000
ИнГок	17000
КМЗ АСЕЛЬМИТАЛ	34000
Общее Число Профессий	7500
По первому списку (очень тяжелые условия труда)	3000
По второму списку (тяжелые условия труда)	5000
Профессиональных поликлиник	12
Исследовательский институт по условиям труда	1

В табл. 1 приведена характеристика количества трудящихся и их профессий Криворожских ГМП. Из данных табл. 1 видно, что ГМП Криворожского бассейна относятся к сложным динамическим и техническим системам, в которых трудится большое количество ГМР. Профессиональные данные их составляют миллиардные записи, которые могут храниться только в хранилищах данных (ХД) [1].

Рассмотрим концепты вопросов, связанные с очисткой, извлечением, преобразованием и загрузкой данных в ХД с позиции подготовки структурного модуля компонента. Исследования в данной области ведутся в различных аспектах и методиках программирования. Широкий диапазон проблематики координирует авторов к спектрализации области для технологии CETL, что позволяет применить предложенные алгоритмы методов CETL для очистки, извлечения, преобразования и загрузки данных мониторинга УТ РМ ГМП. Однако алгоритмизация методов ИТ-технологий имеют закрытый характер и не вписываются в существующие пакеты обработки данных.

3. Цели и задачи исследований

Целью предлагаемой статьи являлось рассмотрение и системный компонент (ориентированный) подхода к применению и использованию алгоритмов технологии CETL [2], упрощающий процесс внедрения их в существующие пакеты обработки данных мониторинга УТ РМ ГМП [2–3].

Для достижения поставленной цели необходимо решить комплекс задач: анализ и преобразование алгоритмов очистки; извлечения; преобразования и загрузки в ХД данных мониторинга УТ РМ ГМП. В результате данного многозадачного подхода реализуется системная компонента анализа для технологического элемента программируемого адаптивного модуля комплексного взаимодействия информационных блоков различного уровня с структурированными этапами ориентированной очистки данных.

4. Решение задачи анализа и преобразования алгоритмов очистки, извлечения, преобразования и загрузки

Процесс идентификации и удаления ошибок, а также несоответствия в данных с целью дальнейшего улучшения их качества называется ОД. Проблемы качества данных встречаются в наборах данных ГМП и необходимость в процессе очистки этих данных существенно возрастает. Это вызвано тем, что источники исходной информации [4] часто содержат неструктурированные данные. Поэтому для организации доступа к идентифицированным и согласованным данным, необходимо обеспечить кластеризацию данных с возможностью исключения дублирующих факторов. В процессе технологии CETL, показанного [5], дальнейшее преобразование данных связано с трансляцией данных и интеграцией их, а также с фильтрацией и агрегацией данных, предназначенных для ХД. Процесс очистки данных выполнялся в отдельной разработанной программе подготовки данных до загрузки преобразованных данных в ХД.

Хранилища данных (ХД) требуют и одновременно обеспечивают всестороннюю поддержку очистки данных, которые загружают и обновляют большие объемы данных из различных источников в реальном режиме, что влечет попадание в них данных с высоким уровнем ошибки или шума. ХД используются для принятия решений ЛПР ГМП, следовательно, чтобы некорректные данные (неструктурированные) не привели к некорректным выводам. Из-за придела спектра несоответствий в данных и большого объема данных, их очистка считается одной из самых крупных проблем в технологии хранилищ данных. Существует множество средств, с различной степенью функциональности, которые предназначены для поддержки смежных задач, зачастую достаточно большой объем работы по очистке, извлечению и преобразованию приходится выполнять вручную или низкоуровневыми программами, трудными для написания и использования оператору или неподготовленному пользователю.

Процессы очистки и интеграции занимают большой временной ресурс мощностей, но позволяет получить полноценные результаты запроса. Метод очистки данных (МОД) ГМП должен удовлетворять ряду критериев [5, 6]: идентификация и удаление основных ошибок и несоответствий в источниках данных, работающих сообща или по отдельности; поддержка специализированного программного инструментария для сокращения объемов ручной проверки и программирования; интеграция колебаний системы и отражения в ее структуре, анализ связей между данными.

Существует множество видов ошибок, которые не зависят от предметной области. Таких ошибок выделяют шесть типов: противоречивость информации; аномальные значения; пропуски

данных; шум; несоответствие форматов данных; ошибки ввода данных или опечатки; дублирование.

Инфраструктура технологического процесса ГМП поддерживается для ХД, обеспечивая эффективное и надежное выполнение всех этапов преобразования данных для множества источников и больших наборов данных. Поскольку большая часть проводимых исследований осуществлено в парадигме трансляции и интеграции данных [4, 7] что имело отражение в едином подходе ОД, касающегося ряда аспектов преобразования, специфических операторов и их реализации.

ОД делят на пять этапов: анализ данных; определение порядка и правил преобразования данных; подтверждение преобразования; противоток очищенных данных. ОД может выполнять одну или несколько функций: парсинг [8] (грамматический или лексический анализ текста и алгоритм процесса это деление полей на атомарные значения); проверка допустимости; стандартизация; согласование и консолидация (практически все средства ОД содержат алгоритмы расстановки приоритетов между полями (в процессе согласования) и контроля очередности сравнения полей); улучшение.

Проведем классификацию основных задач в области МОД ГМП, которые подлежат комплексному решению и связаны между собой. Преобразование данных требуется для поддержки любых изменений в структуре, представлении или содержании данных. В [5] показан процесс разделения функций между задачами со схемой и с элементами данных. Задачи уровня схемы отражаются и в элементах данных; они решаются с помощью ее улучшения, трансляции и интеграции схемы данных. С другой стороны, задачи уровня элемента данных связаны с ошибками и несоответствиями в содержимом текущих данных, незаметных на уровне схемы. Для источников без схемы, существует небольшое количество ограничений на вводимые в ХД данных, что ведет к росту вероятности ошибок и несоответствий [6].

Системы БД характеризуют границы модели данных и их целостности, связанные со спецификой приложения. Проблемы с набором данных, связанные со схемой, происходят из-за недостатка соответствующих моделей или приложений по ограничению целостности. Проблемы, связанные с элементами данных, – результат ошибки или несоответствия, которые невозможно предотвратить на уровне схемы [7]. Ограничения, которые идентифицируются на уровне схемы данных, не защищают систему ОД от дублирования элементов. Анализ схем источников ОД УТ РМ ГМП представляет собой дорогостоящий сложный процесс, поэтому решение задачи уменьшения количества загрязненных данных является определяющим шагом в алгоритме ОД. Для этого спроектирована схема БД УТ РМ ГМП, соответствующим образом, ограничения ее целостности и приложения для ввода данных. Кроме того, выявление правил ОД в процессе проектирования ХД [9, 10], может дать основание для улучшения ограничений, вызванных существующими схемами. Проблемы ОД, представленные отдельными источниками данных, усугубляются в случае интеграции множества источников данных. Каждый источник данных, представленных различным образом, перекрывающих и противоречащих друг другу содержит загрязненные данные. Причиной этому является существование независимых разработок АСУ ГМП, внедрение и поддержка существующих источников, ориентированных на специфические потребности ГМП. В результате проявляется значительная неоднородность в СУД и МД и планах схем оперативных данных, которые, на уровне схем отличия МД и проекта схемы данных, можно обрабатывать в рамках трансляции и интеграции схем данных.

Согласно проведенной классификации МОД [3], средства анализа данных были разделены на средства профайлинга (совокупность психологических методов и методик оценки и прогнозирования поведения горно-металлургического рабочего на основе анализа наиболее информативных частных признаков и характеристик внешности) данных МУТ и средства дэйта майнинг (добычи данных) МУТ [4]. Рассмотрим некоторые существующие математические пакеты [7] (МП) ОД. Так МП «MIGRATIONARCHITECT» решает задачи ОД: тип данных, длину, множество элементов, дискретные значения и их процентное отношение, минимальные и максимальные значения, утраченные значения и уникальность. МП «MIGRATIONARCHITECT» разрабатывает целевые схемы для миграции данных. МП «WIZRULE» и «DATAMININGSUITE» относятся к средствам data mining (добычи данных) выводят отношения между атрибутами и их значениями и вычисляют уровень достоверности, отражающий число квалифицирующих рядов. МП «WIZRULE» может представлять три вида правил: математическую формулу, правила «if-then» и правила правописания, отсеивающие неверно написанные имена, – «WIZRULE» автоматически указывает на отклонения от набора обнаруженных правил так и на возможные ошибки данных. Средства модернизации данных [8], используют обнаруженные шаблоны и правила для определения и выполнения очищающих преобразований, т. е. модернизируют унаследованные данные. Элементы «INTEGRITY» подвергаются воздействию манипуляций обработки – разбору, типизации, анализу шаблонов и частот. Результатом манипуляций является табличное представление содержимого полей, шаблонов и частот, в зависимости стандартизации данных. Для определения очищающих преобразований «INTEGRITY» предлагает язык с набором операторов для преобразований столбцов и рядов. Идентифицирует и консолидирует «INTEGRITY» записи с помощью метода статистического соответствия.

Специализированные средства ОД чаще всего работают в определенном диапазоне [5] – «имена и адреса» с исключением дубликатов. Преобразование

даннях УТ РМ либо обеспечиваются заранее в форме библиотеки правил или в ручном режиме, оператором или пользователем. Преобразование данных МУТ ГМП может быть автоматически получены также с помощью средств согласования схемы. ОД специфической области (имена и адреса записаны в различных источниках) обычно имеют множество элементов данных. Ряд инструментов МП таких как «IDCENTRIC» (FirstLogic), «PUREINTEGRATE» (Oracle), «QUICKADDRESS» (QASSystems), «REUNION» (PitneyBowes) и «TRILLIUM» (TrilliumSoftware) предназначены для очистки именно определенных выше данных УТ ГМП, которые содержат методы по извлечению и преобразованию данных РМ ГМП (таких как «имена и адреса» в отдельные стандартные элементы данных, проверку допустимости названий улиц, городов и индексов, вместе с возможностями сопоставления их на основе очищенных данных). Они [10] включают огромную библиотеку предопределенных правил относительно проблем, часто встречающихся в данных такого рода.

Многие инструменты МП ИТ-технологий поддерживают процесс CETL (очистка, извлечение, трансформация, загрузка) для ХД на комплексном уровне, т. е. единообразное управление метаданными по очищенным источникам данных МУТ, по целевым схемам, маппированию и скриптам ОД, которые они используют в репозитории на основе СУБД МУТ.

Технологический процесс ГМП может также помогать работе внешних средств, например – в специфических задачах очистки, например, таких, как очистка имен/адресов или исключение дубликатов. Средства ETL обычно содержат мало встроенных возможностей очистки, но позволяют пользователю определять функциональность очистки через собственный API. Анализ данных для автоматического выявления ошибок и несоответствий в данных не поддерживается. Оператор может реализовывать такую логику при работе с метаданными и путем определения характеристик содержимого с помощью функций агрегации (sum, count, min, max, median, variance, deviation,…). Поставляемая библиотека преобразований отвечает различным потребностям преобразования и очистки данных – например, конверсию типов данных (в частности, переформатирование данных), строковые функции (например, расщепление, слияние, замена, поиск по подстроке), арифметические, научные и статистические функции и т. д. Идентифицированные [5] оператором функции соответствия полей и функции корреляции сходства полей, программируются и добавляются во внутреннюю библиотеку.

5. Выводы и предложения

Осуществлен анализ с позиции систематизации структурной базы программных компонентов технологии очистки данных. Отображено проекционное формирование программного модуля кросориентированных компонентов SQL и Matlab для проведения исследовательских изысканий по анализу и преобразованию алгоритмов очистки, извлечения, преобразования и загрузки в ХД данных мониторинга УТ РМ ГМП, которые скорректированы в структуре алгоритма программ перехода к существующим математическим пакетам. Исследования подчеркивают необходимость разработки компьютерных промышленных открытых пакетов по методам очистки данных, что может быть результатом дальнейших исследований.

Литература

1. Доровской, В. А. Идентификация профессиональных знаний операторов автоматизированных систем управления [Текст] / В. А. Доровской. – Херсон, 2004. – 354 с.

2. Гаврилова, Т. А. Базы знаний интеллектуальных систем [Текст] / Т. А. Гаврилова, В. Ф. Хорошевский. – Спб.: Питер, 2000. – 384 с.

3. Барсегян, А. А. Анализ данных и процессов: учеб. пособие [Текст] / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д.Тесс, С. И. Елизаров; 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.

4. Введение в Data Mining на базе SQL Server 2008 [Электронный ресурс] / Режим доступа: http://www.techdays.ru/videos/1376.html

5. Эрхард, Рам, Хонг, Хай До Очистка данных: проблемы и актуальные подходы [Электронный ресурс] / Рам Эрхард, Хай До Хонг. – Режим доступа: http://www.iso.ru/rus/document5815.phtml

6. Макленнен, Дж. Microsoft SQL Server 2008: Data mining интеллектуальный анализ данных [Текст] / Дж. Макленнен, Ч. Танг, Богдан Криват; пер. с англ. – СПб.: БХВ-Петербург, 2009. – 720 с.

7. Седова, Н. А. Формирование лингвистических переменных для задач судовождения [Текст] / Н. А. Седова // Эксплуатация морского транспорта. – 2013. – № 2 (72). – С. 19–23.

8. Черный, С. Г. Анализ правил комбинирования групповых экспертных оценок при нечетких данных [Текст] / С. Г. Черный // Системы управления и информационные технологи. – 2014. – № 3.1 (57). – С. 182–187.

9. Чёрный, С. Г. Применение механизма информационных интеллектуальных моделей в системах автоматического управления [Текст] / С. Г. Черный // Вестник Херсонского национального технического университета. – 2012. – № 1 (44). – С. 215–220.

10. Жиленков, А. А. Применение нейронечёткого моделирования для задач идентефикации многокритериальности в транспортной отрасли [Текст] / А. А. Жиленков, С. Г. Черный / Вестник самарского государственного университета путей и сообщений. – 2014. – № 1 (23). – С. 104–110.

References

1. Dorovskoy, V. A. (2004). Identifikatsiya professional'nykh znaniy operatorov avtomatizirovannykh sistem upravleniya. Kherson, 354.

2. Gavrilova, T. A., Khoroshevskiy, V. F. (2000). Bazy znaniy intellektual'nykh sistem. Spb.: Piter, 384.

3. Barsegjan, A. A., Kuprijanov, M. S., Holod, I. I., Tess, M. D., Elizarov, S. I. (2009). Analiz dannyh i processov. SPb.: BHV-Peterburg, 512.

4. Vvedenie v Data Mining na baze SQL Server 2008. Available at: http://www.techdays.ru/videos/1376.html

5. Jerhard, Ram, Hong, Haj Do Ochistka dannyh: problemy i aktual'nye podhody. Available at: http://www.iso.ru/rus/document5815.phtml

6. Maklennen, Dz., Chzhaohujej, T., Krivat, B. (2009). Microsoft SQL Server 2008: Data mining intellektual'nyj analiz dannyh. SPb.: BHV-Peterburg, 720.

7. Sedova, N. A. (2013). Formirovanie lingvisticheskih peremennyh dlja zadach sudovozhdenija. Jekspluatacija morskogo transporta, 2 (72), 19–23.

8. Chernyi, S. G. (2014). Analiz pravil kombinirovanija gruppovyh jekspertnyh ocenok pri nechetkih dannyh. Sistemy upravlenija i informacionnye tehnologii, 3.1 (57), 182–187.

9. Chernyi, S. G. (2012). Primenenie mehanizma informacionnyh intellektual'nyh modelej v sistemah avtomaticheskogo upravlenij. Vestnik Hersonskogo nacional'nogo tehnicheskogo universiteta, 1 (44), 215–220.

10. Zhilenkov, A. A., Chernyi, S. G. (2014). Primenenie nejronechjotkogo modelirovanija dlja zadach identefikacii mnogokriterial'nosti v transportnoj otrasli. Vestnik samarskogo gosudarstvennogo universiteta putej i soobshhenij, 1 (23), 104–110.

Дата надходження рукопису 30.10.2014

Доровской Владимир Алексеевич, доктор технических наук, профессор, кафедра электрооборудования судов и автоматизации производства, Керченский государственный морской технологический университет, ул. Орджоникидзе, 82, г. Керчь
E-mail: dora_1943@mail.ru
Черный Сергей Григорьевич. кандидат технических наук, доцент, кафедра электрооборудования судов и автоматизации производства, Керченский государственный морской технологический университет, ул. Орджоникидзе, 82, г. Керчь
E-mail: sergiiblack@gmail.com
Доровская Ирина Александровна, аспірант, кафедра информационных технологий, Херсонский национальний технический університет, ул. Бориславсое шоссе, 24, г. Херсон, Украина, 28300
Сметюх Надежда Павловна, аспирант, кафедра электрооборудования судов и автоматизации производства, Керченский государственный морской технологический университет, ул. Орджоникидзе, 82, г. Керчь
E-mail: golosa @mail.ru

УДК 629.7.054
DOI: 10.15587/2313-8416.2014.29127

РЕЗОНАНС СПІВПАДІННЯ І ПОХИБКИ ПОПЛАВКОВОГО ГІРОСКОПА

© В. В. Карачун, В. Ю. Шибецький

Наведена модель явища появи додаткових похибок при взаємодії N-хвилі з приладами інерціального позиціонування гіперзвукових літальних апаратів. Розраховані кути співпадіння для датчика кутової швидкості. Дані розрахунків порівняні з експериментальними даними, отриманими при напівнатурних стендових дослідженнях для двох випадків, з працюючим і вимкненим гіроагрегатом. Зроблені висновки стосовно можливості використання отриманих результатів.
Ключові слова: гіперзвукові літальні апарати, N-хвиля, зона каустик, хвильове співпадіння, поплавковий гіроскоп.

The model of effects of additional errors in the interaction of N-wave with inertial positioning devices of hypersonic aircraft was created. The angles of coincidence to the angular velocity sensor were calculated. These calculations are compared with experimental data obtained in research of semi natural conditions for two cases of running and non-running gyro unit. Conclusions of possibility of results usage were made.
Keywords: hypersonic aircraft, N-wave, caustic zone, wave coincidence, floating gyroscope.

1. Вступ

За висновками аналітиків, в найближчому майбутньому гіперзвукові технології будуть в змозі забезпечити захист стратегічних інтересів будь-якої країни, в світі.

Значення *гіперзвукових технологій* неможливо переоцінити. Бойовий гіперзвуковий літальний апарат (ЛА) отримує величезну тактичну перевагу над існуючими бойовими засобами, стаючи практично неуразливим для сучасних систем ППО [1]. Вже при швидкості 4 *M* гіперзвукова крилата ракета здатна перелетіти Атлантичний океан менш ніж за годину.

До недавнього часу, розробку і випробування гіперзвукових ЛА здійснювали лише дві країни – США і Росія. Відтепер, до перших двох приєдналася третя країна – Китай, із своїми гіперзвуковими технологіями. Випробування апарату, за твердженням китайських джерел, несе виключно наукове значення. Однак, в перспективі, на основі даної конструкції можуть бути створені крилаті ракети [2].

Незважаючи на те, що Китай не має такої тривалої історії створення гіперзвукових ЛА як США і Росія, проте його літальний апарат, за даними засобів масової інформації, вже долає швидкість у 10 *M*, і це при тому, що подібні ЛА в Росії рухаються зі швидкістю лише в 4,5 *M*, а в США – 5 *M* [3].