УДК 65.011.56

Применение элементов модели интеллектуального анализа данных для оценки работоспособности морских буровых комплексов

С.Г. Черный
к.т.н., доцент[1]
sergiiblack@gmail.com

[1]Керченский государственный морской технологический университет, Керчь, Россия

Исследована и разработана модель интеллектуального анализа данных условий рабочих мест оператора морских буровых платформ. Осуществлен анализ современных информационных продуктов и надстроек, которые используются в данной области. Проанализированы методы и средства очистки данных морских буровых платформ в разрабатываемых информационных системах. Представлен наглядный процесс взаимодействия и надстроек в SQL Server и Management Studio в качестве элемента хранения и представления данных.

Материалы и методы
На основе технологии оптимизации работы потоковых данных и обработки информации даны рекомендации и сформулированы методы оптимизации взаимодействия составных компонентов морских буровых платформ.

Ключевые слова
хранилище прецедентов, база данных, алгоритм, буровая платформа

На сегодняшний день в корпорации, какими являются морские добывающие платформы, в процессе бурения, разведки и эксплуатации поступает и обрабатывается большой объем данных, особенно технологических и персональных, собранных со всех систем. В каждом комплексе таких «систем» реализована своя структура базы данных, и после интеграции данных в единый источник — хранилища данных (ХД), возникает проблема извлечения достоверных данных по причине их разрозненности в различном представлении, которые необходимо в дальнейшем использовать для анализа. Такие данные будут низкого качества, так как в них допускались ошибки, и обрабатывать их теряет всякий смысл. Поэтому для получения реальных выводов из существующих данных применяют различные методы по их коррекции, исключению дубликатов и очистки.

Решение задачи

Известны три причины возникновения ошибок в данных:

1. в основном сведения вводятся операторами вручную (т.к. присутствует фактор невнимательности, то допускаются опечатки в словах, не заполняются обязательные поля при анкетировании, сокращаются названия, заносятся сведения не в те поля и т.д.);
2. не во всех существующих программах и модулях надстройки, в которые вносится информация, указаны элементы ограничения для их значения;
3. в МБП сбор информации о технических параметрах ведется несколькими комплексами, следовательно, при слиянии всех этих сведений в единую базу данных возникают проблемы с форматами однотипных данных.

Существует большое количество видов ошибок, которые не зависят от характера предметной области. Подобного рода ошибки выделяют в семь индикаторов:

- противоречивость информации;
- аномальные значения;
- пропуски данных;
- шум;
- несоответствие форматов данных;
- ошибки ввода данных или опечатки;
- дублирование.

Противоречивостью информации называют такую информацию, которая не соответствует законам, правилам или действительности. Сначала решается, какую именно информацию необходимо считать противоречивой. Такие значения параметров информации корректируют вручную, что связано с проектированием программных средств прогнозирования, которые не учитывают природу процессов прогнозирования, что характерно будет восприниматься (аномальная группа), как совершенно нормальное значение, и сильно искажать картину будущего прогноза, т.е. случайный провал или успех будет считаться закономерностью [2, 4]. Пропуски данных — тип ошибок, когда в полях для заполнения отсутствуют или заполнены данными не до конца и такого рода проблема очень серьезная для большинства ХД.

Большое количество методов прогнозирования исходят из предположения, что данные поступают равномерным постоянным потоком. На практике такое поступление данных встречается редко. Поэтому одна из самых востребованных областей применения ХД — прогнозирование — оказывается реализованной некачественно или со значительными ограничениями.

Достаточно часто в процессе анализа информации с комплексов датчиков, сталкиваются с шумами, которые не несут никакой информации и мешают четко идентифицировать состояние процесса. Ошибки ввода данных или некорректное их отображение преобладают в любых данных, т.к. вводятся оператором.

Рассмотрим этапы очищения данных. Очистку данных делят на пять этапов: анализ данных; определение порядка и правил преобразования данных; подтверждение; преобразования; противоток очищенных данных.

На начальной итерации детализировано осуществляется анализ данных для выявления подлежащих к удалению видов ошибок и неточностей. Зачастую, используются два вида проверок данных: вручную или специальными программами. На этом этапе осуществляется процесс получения метаданных о свойствах и их качестве, а далее идентифицируется порядок и правила преобразования данных. В зависимости от набора источников данных, степени их неоднородности и загрязненности, могут быть преобразованы.

Для отображения источников обобщенной модели данных на практике иногда используется трансляция схемы, а для ХД — реляционное представление. Начальные шаги для очистки данных направлены на идентификацию проблем отдельных источников данных. Дальнейшие шаги должны быть направлены на интеграцию схемы данных и устранение проблем участков множественных элементов (дубликатов). Для ХД в процессе работы по определению ETL (Extract, Transform, Load — дословно «извлечение, преобразование, загрузка» [1]) должны быть определены методы контроля и поток данных, подлежащий преобразованию и очистке. На третьей итерации определяются два атрибута очистки данных: правильность и эффективность процесса, и определение преобразования. Процесс осуществляется путем тестирования и оценивания технических данных МБП. При анализе, проектировании и подтверждении

данных МБП может потребоваться множество итераций, в связи с тем, что некоторые ошибки становятся заметны только после проведения определенных преобразований данных МБП. На четвертом — осуществляется процесс выполнения преобразований или ETL для загрузки и обновления ХД, или при ответе на запросы по множеству источников. На пятом этапе происходит замена загрязненных данных в исходных источниках на очищенные. Это необходимо осуществить для того, чтобы улучшенные данные МБП попали также в унаследованные приложения и в дальнейшем при извлечении не требовали дополнительной очистки. Для хранилищ очищенные данные находятся в области хранения данных [1].

Анализируя методы и средства очистки данных МБП в разрабатываемых информационных системах, на сегодняшний день существует значительное количество методов для очистки данных от ошибок и неточностей. Эта группа методов разносторонне подходит к проблеме очистки данных от ошибок и неточностей, а проблема, данного рода решается тремя способами: простыми методами; методами, которые основываются на понятиях математической статистики; средства ETL.

Простые методы (регулярные выражения, строгие формальные правила и т.д.) решают данную задачу только частично, и поэтому в работе использовали методы математической статистики. Осуществляется расчет необходимых групп факторов по всем данным, которые есть в наличии, т.е. охватывается весь диапазон значений и принимаемых признаков. На основе полученных результатов одни методы могут выделить информацию, которая сильно отличается от остальных, а другая группа методов — вычисляет величины, которые предположительно более всего похожи на истинные. Осуществляя анализ сведений с помощью статистических характеристик, оценивают общую картину данных, и на фоне определяют возможные ошибки с последующими их корректировками на подобранные похожие значения. В работе использовались такие методы очистки данных:

- Устранялись типы ошибок (аномалии, пропуски, неправдоподобие данных и опечатки). В данном методе осуществлены подсчеты частоты появления определенного значения в имеющихся данных. На начальных итерациях суммировано, какое количество раз различные значения были введены. Затем сортировались их частоты по убыванию. В итоге, в конце списка находятся значения, которые реже всего пользователь вводил. Можно предположить, что в данных допускались опечатки, наведены значения или введены аномальные значения. Поэтому такие поля подвергались дополнительной обработке и последующей замене. После обнаружения данных с низким качеством, использовался простой метод — анализ строк, с помощью него восстанавливают вероятные значения.
- Вычисление средних значений для устранения пропусков. Вычисляется 3 типа значений: мода, медиана и среднее арифметическое значение. При условии, что данные содержат большой разброс значений, метод средних применяется не к отдельному объекту, а к целой группе. Наборы данных в этом случае разбиваются на группы, которые содержат приблизительно однородные элементы с похожими признаками. Внутри каждой из групп (наборов) рассчитывалась средняя величина входящих в данный кортеж (группу).
- Интервальный метод используется при условии, что набор (выборка) данных является не разнородными. Данным методом производится расчет доверительного интервала, между границами которого с заданной вероятностью находятся истинные значения оцениваемых параметров. Доверительный интервал с вероятностью 95% для большого объема данных, подчиняющихся нормальному закону распределения, определяют по формуле:

$$\bar{x} - \frac{1.96x\sigma}{\sqrt{n}} < x_i < \bar{x} + \frac{1.96x\sigma}{\sqrt{n}} \qquad (1)$$

где x_i — исследуемый ряд данных, x — среднее арифметическое значение совокупности данных, σ — среднеквадратическое отклонение, n — количество исследуемых данных.

Значения, не попавшие в этот интервал, отмечаются как потенциальные ошибки, их заменяют уже подобранными значениями (например, средней арифметической величиной). Метод применялся для однородных данных.

Третий способ решения задачи базируется на использовании ETL средств для ХД. ETL средства включают в себя три основных процесса: извлечение данных из внешних источников; преобразование данных и их очистка; загрузка в ХД [2]. Такие средства обеспечивали возможность сложных преобразований и большей части технологического процесса преобразования и очистки данных. Общей проблемой средств ETL являются ограниченные за счет собственных API и форматов метаданных возможности взаимодействия, усложняющие совместное использование различных средств. На многих МБП инструменты поддерживают процесс ETL для ХД на комплексном уровне. Для единообразного управления всеми метаданными по источникам данных, целевым схемам, манипулированием, скриптам и т.д. они используют репозитарий на основе СУБД. Схемы и данные извлекались из оперативных источников данных как через свой файл и шлюзы СУБД DBMS, так и через стандартные интерфейсы — например, ODBC и EDA. Преобразование данных осуществляется через простой графический интерфейс. Для определения индивидуальных итераций шагов (маппирования) чаще всего используют собственный язык правил и набор библиотек предопределенных функций преобразований. Данный набор средств поддерживает и повторное использование существующих преобразованных решений, например внешних процедур C/C++ с помощью имеющегося в них интерфейса для их интеграции во внутреннюю библиотеку преобразований. Процесс преобразования выполняется системой, интерпретирующей специфические преобразования в процессе работы или откомпилированным кодом. Все средства на базе системы имеют планировщик и поддерживают технологические процессы со сложными зависимостями выполнения между этапами преобразования. Технологический процесс ОД оказывает поддержку в работе внешних средств, а инструментарий ETL [3] применяют для очищения персональных данных. Преобразование осуществляется двумя способами: в форме библиотеки правил заранее или оператором в интерактивном режиме. Отметим, что данные могут быть автоматически получены и с помощью средств согласования схемы. За счет ограниченной области, такие средства обычно очень эффективны, но не лишены недостатков: нуждаются в надстройках другими инструментами для работы с широким спектром проблем преобразования и очистки. Очистка данных может выполнять одну или несколько функций. Например, парсинг (грамматический или лексический анализ текста и в процессе его выполнения осуществляется деление полей на атомарные значения). Некоторые приложения объединяются с такими программами, с помощью которых можно сверить данные.

Компоненты DATACLEANSER и MERGE/PURGE LIBRARY позволяют интегрировать команды/правила согласования, определенные пользователем АСДД [1]. При существовании множества платформ, систем, инструментов для преобразования и очистки данных, идеального решения нет, и все они не избавляют от процесса дублирования,

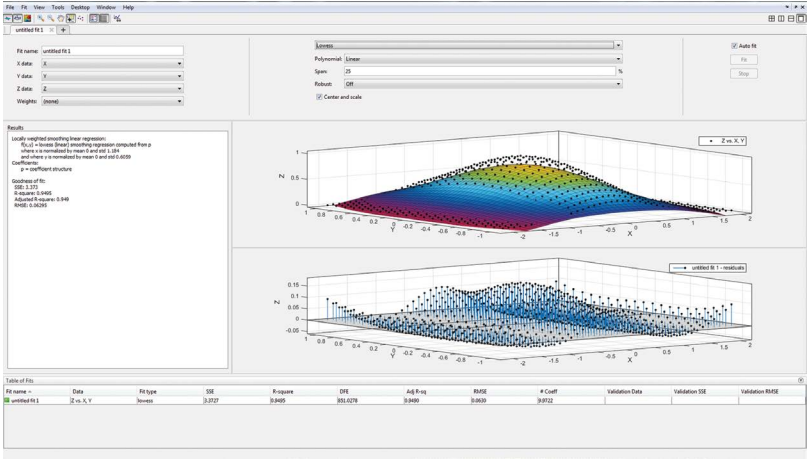

Рис. 1 — Пространственное представление области очищенных данных МБП в Curve Fitting Toolbox

потери данных и несоответствия. Поэтому и сейчас специалисты пытаются найти оптимальные вариации для решения очистки данных. Информационные потоки данных чаще всего реализованы в виде таблиц, которые могут быть представлены как одноуровневыми так и многоуровневыми слоями структур [5, 6]. При дальнейшем анализе результатов именно формат электронной таблицы позволяет использовать пакеты анализа Matlab. Алгоритм в математическом пакете MatLab (Image Toolbox & Curve Fitting Toolbox) можно представить диаграммой для удобства визуализации информации. Анализ результатов проведен с использованием Curve Fitting Toolbox пакета MATLAB.

Результаты анализа данных представлены в виде таблицы, в которой названия всех требуемых параметров представлены в правой части таблицы, а их числовое значение — в левой. Этот анализ позволил построить диаграмму пространственного представления очищенных данных МБП. Пространственное представление показано на скриншоте (рис. 1) в Curve Fitting Toolbox.

Развитие современного программного обеспечения в отрасли нефтегазового сектора с элементами интеллектуального анализа достаточно перспективно на территории Крымского полуострова. Происходит переход на стандарты технологических карт, оптимизации и настройки аппаратно-программных компонентов. Создание собственных программных и независимых комплексов АСУ нефтегазового сектора в России для отрасли является приоритетным и только набирает рост, но уступает разработкам зарубежных стран. Большинство буровых платформ нефтегазового сектора снабжены иностранным программным обеспечением закрытого архитектурного типа.

Итоги

Приведены методика расчёта и технология обработки данных на морских буровых платформах с учетом специализированного программного обеспечения. Дано обоснование по применению и оптимизации параметров систем контроля за технологическим циклом.

Выводы

Установлено, что модель интеллектуального анализа данных МБП перед обработкой структуры является просто контейнером, который задает столбцы, используемые для входных данных, прогнозируемый атрибут и параметры, управляющие алгоритмом обработки данных. В процессе санкций и импортозамещения, развитие современных программных технологий с элементами интеллектуального анализа достаточно перспективно. Согласно конфигурации части АСУ БП, на Крымском полуострове предусмотрена возможность дистанционного контроля и управления, в том числе и через спутниковую связь из офиса службы поддержки «National Oilwell Varco».

Список используемой литературы

1. Макленнен Дж., Чжаохуэй Т., Криват Б. Microsoft SQL Server 2008: Data mining — интеллектуальный анализ данных. СПб.: БХВ-Петербург, 2009. 720 с.
2. Барсегян А. А., Куприянов М.С., Холод И.И., Тесс М.Д, Елизаров С.И. Анализ данных и процессов: учеб. пособие 3-е изд. СПб.: БХВ-Петербург, 2009. 512 с.
3. Службы SQL Server Analysis Services – интеллектуальный анализ данных. Режим доступа: http://msdn.microsoft.com/ruru/library/bb510517.aspx (дата обращения 19.10.2015).
4. Черный С.Г., Жиленков А.А. Оценка надежности функционирования морских буровых платформ // Автоматизация, телемеханизация и связь в нефтяной промышленности. 2015. № 1. С. 30–36.
5. Chernyi S.G. The problems of automation technological process of drilling oil and gas wells // Программные продукты и системы. 2015. №2. С. 113–118.
6. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. New York: Wiley, 2004, 528 p.

ENGLISH **AUTOMATION**

Application of elements models intellectual data analysis for estimation of serviceability of maritime drilling complexes

UDC 65.011.56

Authors:
Sergei G. Chernyi — Ph.D., associate professor[1]; sergiiblack@gmail.com

[1]Kerch State Maritime Technological University, Kerch, Russian Federation

Abstract

It was researched and developed the mining model conditions of operator offshore drilling platforms. The analysis of modern information products and add-ons, which are used in the field. The analysis of methods and tools for data cleansing offshore drilling platforms in the emerging information systems. It is presented a clear process of communication and add-in SQL Server Management Studio, and as part of the storage and presentation of data.

Materials and methods

Based on the optimization of the technology of streaming data and processing information and recommendations formulated methods of optimizing interaction between the components of the offshore drilling platforms.

Results

The methodology of calculation and data processing technology on offshore platforms based on specialized software. The substantiation of the application and optimization of parameters of control systems for technological cycle.

Conclusions

It was established that the mining model ODP before processing structure is just a container that specifies the columns used for input, and predictable attribute settings that control the data processing algorithms. During the sanctions and import substitution, the development of modern software technology with elements of mining quite promising. According to the configuration of the ASC ODP on the Crimean peninsula provides the possibility of remote monitoring and control, including via satellite office support service «National Oilwell Varco».

Keywords

storage cases, database, algorithm, drilling rig

References

1. Maklennen Dzh., Chzhaokhuey T., Krivat B. *Microsoft SQL Server 2008: Data mining — intellektual'nyy analiz dannykh* [Microsoft SQL Server 2008: Data mining]. St. Petersburg: *BKhV-Peterburg*, 2009,720 p.
2. Barsegyan A. A., Kupriyanov M.S., Kholod I.I., Tess M.D, Elizarov S.I. *Analiz dannykh i protsessov* [Analysis of data and processes]. Study book, 3-rd ed. St. Petersburg: *BKhV-Peterburg*, 2009, 512 p.
3. *Sluzhby SQL Server Analysis Services – intellektual'nyy analiz dannykh* [SQL Server Analysis Services —data mining]. Available at: http://msdn.microsoft.com/ruru/library/bb510517.aspx (accessed 19 October 2015).
4 Chernyy S.G., Zhilenkov A.A. *Otsenka nadezhnosti funktsionirovaniya morskikh burovykh platform* [Serviceability of maritime drilling complexes]. *Avtomatizatsiya, telemekhanizatsiya i svyaz' v neftyanoy promyshlennosti*, 2015, issue 1, pp. 30–36.
5. Chernyi S.G. The problems of automation technological process of drilling oil and gas wells. *Programmnyye produkty i sistemy*, 2015, issue 2, pp. 113–118.
6. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. New York: Wiley, 2004, 528 p.