ОБРАБОТКА СИГНАЛОВ И ДАННЫХ, РАСПОЗНАВАНИЕ ОБРАЗОВ, ВЫЯВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ И ПРОГНОЗИРОВАНИЕ

УДК 44.01.81

ПРОБЛЕМЫ КАЧЕСТВА ДАННЫХ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ КОММЕРЧЕСКОГО УЧЕТА ПОТРЕБЛЕНИЯ ЭНЕРГОРЕСУРСОВ

Статья поступила в редакцию 17.04.2014, в окончательном варианте 23.04.2014.

Савкина Анастасия Васильевна, кандидат технических наук, доцент, Мордовский государственный университет, 430005, Российская Федерация, г. Саранск, ул. Б. Хмельницкого, 39, e-mail: av-savkina@yandex.ru

Федосин Александр Сергеевич, аспирант, Мордовский государственный университет, 430005, Российская Федерация, г. Саранск, ул. Б. Хмельницкого, 39, nsdfxela@gmail.com

В статье рассматриваются основные вопросы качества данных, используемых в автоматизированных системах начислений платы за потребленные энергоресурсы. Согласно общепринятой классификации определяются основные свойства источников данных, а также связанные с ними проблемы. Предлагается способ выявления некорректных показаний приборов учета, основанный на статистической оценке средних расходов. Также описывается возможность использования коэффициента асимметрии для определения подозрительных (сомнительных) значений. Кроме того, в качестве одного из инструментов решения задачи рассмотрено использование коэффициента эксцесса. Определение коэффициента асимметрии представлено как часть ETL процесса, а среди предполагаемых программных библиотек и средств его реализации для платформы .NET выделены те, которые, по мнению авторов статьи, содержат минимально необходимый функционал. Отдельное внимание уделяется потенциальным проблемам интеграции биллинговых систем с автоматизированными системами коммерческого учета энергоресурсов (АСКУЭ) в свете проблематики качества данных. Описываются механизмы, позволяющие использовать метаданные статистического характера, для совершенствования процесса соотнесения сведений о потреблении энергоресурсов в процессе интеграции.

Ключевые слова: качество данных, профилирование данных, коммерческий учет потребления энергоресурсов, АСКУЭ, коэффициент асимметрии, коэффициент эксцесса, ETL-процесс, хранилища данных, интеграция данных, метаданные

DATA QUALITY ISSUES FOR ENERGY MANAGEMENT SYSTEMS

Savkina Anastasiya V., Ph.D. (Engineering), Associate Professor, Mordovia State University, 39 B. Hmelnitsky St., Saransk, 430005, Russian Federation, e-mail: av-savkina@yandex.ru

Fedosin Alexander S., post-graduate student, Mordovia State University, 39 B. Hmelnitsky St., Saransk, 430005, Russian Federation, e-mail: nsdfxela@gmail.com

The article presents key aspects of data quality issues in utility billing management systems. We classify data sources specific properties and related problems. The paper also provides an approach towards incorrect meter readings, which is based on statistical analysis of average resource consumptions. We measure skewness with respect to normal distribution of readings to figure out if there are suspicious values. Kursis is also considered as potential part of outlier detection mechanism. Moreover, we describe the skewness measurement as a part of ETL process and provide list of tools which might be used for such process devel-

opment in .NET framework and implement a required minimum functionality. With regard to SCADA (supervisory control and data acquisition) systems key data quality problems are also noticed. Then a discussion on integration problems and its possible solutions follows. We also describe an idea of using the statistical metadata for mapping performance improving.

Keywords: data quality, data profiling, energy resources accounting, SCADA, scewness, kursis, ETL-process, data warehouse, data integration, metadata

Согласно высказыванию известного специалиста в области искусственного интеллекта Эндрю Нг, при решении задач в области машинного обучения «побеждает не тот, кто обладает лучшими алгоритмами, а тот, у кого больше данных» (It's not who has the best algorithm that wins. It's who has the most data) [10]. Однако работа с большими объемами информации на практике, помимо богатых возможностей для аналитики и выявления знаний, зачастую приносит множество проблем, причиной которых является низкое качество данных. Несмотря на то, что вопросам анализа и управления качеством данных посвящен ряд работ [4, 7, 8], отдельные аспекты этой проблематики остаются раскрытыми недостаточно полно. Поэтому целями данной работы были следующие:

а) определить круг проблем качества данных, общих и специфических для исследуемой предметной области;

б) классифицировать проблемы и возможные решения в соответствии с общепринятыми способами;

в) предложить решения и описать математический аппарат, который может быть использован в них;

г) предложить конкретные технические средства реализации описанных решений.

В общем случае проблемы качества данных в информационных системах классифицируют следующим образом [8]:

1) в зависимости от возможности решить проблему качества добавлением новых ограничений в схеме БД (уровень схемы). Случай, при котором добавление новых бизнес-правил не позволит повысить качество данных (instance level), требует более сложных методов решения;

2) по количеству источников, из которых данные поступают в хранилище или базу данных. Проблемы, связанные с интеграцией данных из нескольких источников, имеют свою специфику по сравнению с теми, которые возникают при получении информации из одного.

В контексте крупных биллинговых систем проблемы, связанные с некорректностью входной информации, в итоге могут становиться причиной выставления неправильных счетов клиентам. Одной из областей, в которых такие ошибки представляются наиболее ощутимыми, является сфера расчетов за объемы потребленных энергоресурсов в жилищно-коммунальном хозяйстве. Острой эту проблему делают следующие факторы:

1) наличие большого числа сторон, заинтересованных в контроле за результатами расчетов (поставщики энергоресурсов, управляющие компании, надзирающие органы, непосредственно жильцы многоквартирных домов);

2) масштабность процессов и сложность обработки данных (в крупных городах начисления могут производиться для сотен тысяч лицевых счетов, причем каждый из них может быть связан с несколькими приборами учета);

3) наличие повышенного интереса к проблемам отрасли в обществе.

Расчет потребленных энергоресурсов зависит от двух факторов [6]:

1) установлен ли коллективный (общедомовой) прибор учета для той или иной коммунальной услуги;

2) установлены ли индивидуальные или общие (квартирные) приборы учета в жилых (нежилых) помещениях в данном доме.

Общая тенденция роста числа установленных приборов учета в многоквартирных домах в конечном итоге приводит к увеличению масштабов поступающей информации об объемах потребленных коммунальных ресурсов [3]. Такая информация может поступать из различных источников: например, показания электросчетчиков, установленных в квартире, могут быть переданы непосредственно жильцами, установлены проверяющим лицом, или агрегированы при помощи автоматизированных системамх коммерческого учета энергоресурсов (АСКУЭ) (в зарубежной терминологии – SCADA систем). В таких условиях имеется достаточно высокая вероятность возникновения проблем качества данных. Они особенно характерны для хранилищ, объединяющих информацию из различных источников. Причем в большинстве своем эти проблемы не могут быть решены внесением изменений в схему баз данных (БД). Можно привести следующие примеры:

- несогласованность данных из разных источников – например, в случае, если контролер фиксирует расход выше того, о котором сообщают жильцы;
- ошибки интеграции систем АСКУЭ, связанные с тем, что по тем или иным причинам к расчету принимаются показания прибора, установленного в неправильном месте, или же, для многотарифных приборов используются значения, определенные по несоответствующему тарифу (дневной вместо ночного и наоборот);
- получение сведений об объеме потребленного ресурса вместо значений расхода. Объем в этом контексте представляет собой разницу между значениями (показаниями приборов учета), установленными в двух расчетных периодах.

Главным и важнейшим этапом процесса очистки данных (data cleaning) является их анализ. Этот этап подразумевает выявление ошибок и пропущенных данных, причем самым распространенным и очевидным способом решения этих задач является ручная обработка. Однако, исходя из описанных выше особенностей предметной области, для эффективной аналитической работы необходимы специальные программные решения.

Одним из известных подходов является профилирование данных (data profiling), которое подразумевает под собой анализ отдельных атрибутов сущностей. Основой для такого анализа служат метаданные, такие как: тип данных, длина, диапазон допустимых значений и т.д.

Далее представлен способ профилирования данных о потреблении энергоресурсов, основанный на статистических показателях полученных значений.

Для работы с имеющимися данными необходимо учесть следующую особенность: сведения о потреблении ресурса поступают с нерегулярной периодичностью (в типичных случаях период может отличаться от календарного месячного в пределах 4–5 дней). Кроме того, существуют ситуации, при которых показания прибора учета могут не передаваться на протяжении длительного времени. Таким образом, данные для анализа имеет смысл представить в следующем виде:

$$x_i = L/(D_{i+1} - D_i), \ ..\text{i} = 1, 2, 3,...\text{n} \qquad (1)$$

где n – количество анализируемых показаний конкретного прибора учета; L – объем потребленного ресурса; $(D_{i+1} - D_{i-1})$ представляет собой разницу между датами получения показаний, на основе которых определен расход L.

Процесс выявления аномальных показаний приборов учета базируется на следующем алгоритме.

1. Выбирается определенный закон распределения случайной величины.

2. Исходя из предположения, что все значения анализируемого ряда подчиняются выбранному распределению, вычисляются параметры (среднее, стандартное отклонение).

3. Аномалиями считаются показания, для которых низка вероятность быть сгенерированными распределением с рассчитанными параметрами (для которых значение не попадает в пределы ±3σ) [9].

Пусть $x \in R$. Если принять, что

$$\hat{x} \sim N(\mu,\ \sigma^2), \tag{2}$$

т.е. величина распределена по нормальному закону, то зная, что данное распределение параметризуется через медиану

$$\mu = (1/n)\sum_{i=1}^{n} x_i \tag{3}$$

и дисперсию

$$\sigma^2 = (1/\mathrm{n})\sum_{\mathrm{i}=1}^{\mathrm{n}} (x_i - \mu)^2\,, \tag{4}$$

можем определить значение $\rho(x)$ для рассматриваемого случая:

$$\rho(x) = \rho(x_j;\mu_j,\sigma^2)(1/\sqrt{2\pi\sigma_j})\exp(-(x_j-\sigma_j)^2/(2\sigma_j^2)). \tag{5}$$

Аномальными будем считать такие значения, для которых $\rho(x) < e$, где $e \le 1$ – некоторое пороговое значение, с помощью которого можно регулировать работу алгоритма. Те значения, которые в результате работы алгоритма будут признаны аномальными, должны получить статус «контрольно-информационных». Это прежде всего означает, что они, в отличие от «контрольных» показаний, не будут приняты к расчету. Кроме того, аномальные показания могут быть визуально отображены с помощью отчетных систем хранилищ данных. Такие отчеты в дальнейшем должны быть использованы для анализа причин появления «сомнительных» данных.

Как было отмечено выше, среди проблем качества данных о расходах энергоресурсов в жилых многоквартирных домах можно выделить те, которые связаны с интеграцией данных АСКУЭ. Эти проблемы обычно связаны с нарушениями в логике коммутации передающих подсистем. В таком случае показания одного передающего устройства ошибочно принимаются за показания другого, или же, к учету показаний принимаются измерения иного ресурса. Например, теплосчетчик-регистратор марки «Взлет ТСР-024М» может одновременно использоваться для измерения потребляемого количества тепла, холодной и горячей воды в трех теплосистемах. В случае использования для расчетов показаний большого числа подобных приборов появляется риск использования неверных данных.

Для решения подобных проблем интеграции данных тоже могут использоваться описанные выше статистические параметры. Алгоритм контроля ошибок интеграции можно представить следующим образом (рис. 1).

Имея статистические показатели, сохраненные в таблицах БД, с помощью возможностей языка запросов можно разделить их на N равных групп [5] (рис. 2).

Вход: множество W приборов учета в хранилище, множество S приборов учета системы АСКУЭ, множества показаний соответствующих приборов I_s, I_w.

```
minSDate := min(I_s.ValueDate);
maxSDate := max(I_s.ValueDate);
для всех w ∈ W
      mean[w] := Расчет-Среднего(I_w[w], minSDate, maxSDate);
      dispersion[w] := Расчет-Дисперсии(I_w[w], minSDate, maxSDate);
      Добавить-Статистику-Счетчика-БД (w.id, mean[w], dispersion[w]);
end;
для всех s ∈ S
      mean[s] := Расчет-Среднего(I_s[s], minSDate, maxSDate);
      dispersion[s] := Расчет-Дисперсии(I_s[s], minSDate, maxSDate);
      Добавить-Статистику-АСКУЭ-БД (s.id, mean[s], dispersion[s]);
end;
```

Рис. 1. Алгоритм контроля ошибок интеграции, записанный в псевдокоде

```
;WITH NTILES AS
(
SELECT NTILE(@N) OVER (ORDER BY cs.Dispersion) as N
                ,cs.Dispersion
                ,cs.CounterId
FROM              dbo.CounterStatistics  cs
)
UPDATE cs1
SET cs1.Ntile = NTILES.N
FROM NTILES INNER JOIN dbo.CounterStatistics  cs1
```

Рис. 2. SQL-запрос, присваивающий номера групп статистическим показателям

Теперь помимо прочих атрибутов (данных о местоположении прибора учета, типе измеряемого ресурса, разрядности и т.д.), при интеграции данных из сторонних систем можно пользоваться статистическими оценками показаний счетчиков. Для этого можно оценивать «попадание» объекта из внешней системы в ту или иную группу на основе рассчитанных показателей.

Справедливым является вопрос правомерности выбора нормального распределения для анализа показаний счетчиков. В связи с этим следует отметить, что особенности решаемой задачи накладывают ограничения на инструментарий: количество хранимых показаний для приборов учета на практике редко превышает несколько десятков значений, а значит, проверка нормальности распределения может быть проведена лишь достаточно приближенно. В качестве механизма проверки нормального характера распределения для небольших объемов выборок ($n \leq 25$) может быть применен способ, основанный на определении асимметрии и отношения средних отклонений, а также использовании специальных таблиц [2].

Прежде всего, необходимо определить значение коэффициента асимметрии β, рассчитанного как отношение центрального момента третьего порядка (μ^3) к среднеквадратическому отклонению в 3-й степени (σ^3):

$$\beta = \mu^3/\sigma^3, \tag{6}$$

$$\beta = (M(x - Mx)^3)/\left[M(x - Mx)^2\right]^{3/2}. \tag{7}$$

Далее определяется величина d_n средней абсолютной ошибки:

$$d_n = \left((1/\mathrm{n}) \sum_{\mathrm{i}=1}^{\mathrm{n}} \left| x - Mx \right| \right) / \left((\sqrt{(1/\mathrm{n})} \sum_{i=1}^{n} (x - Mx)^2 \right). \quad (8)$$

С использованием табличных значений определяются величины процентных точек $\gamma_{1.Q_1}$, $d_{n.Q_2}$, $d_{n.1-Q_2}$ [2]. Если хотя бы одно из неравенств

$$\beta_1(n) < \gamma_{1.Q_1}(n);\ d_{n.1-Q_2}, < d_n < d_{n.Q_2} \quad (9)$$

оказалось нарушенным, то гипотеза нормальности отвергается с уровнем значимости α, подчиняющимся неравенствам [1]:

$$2max(Q_1,\ Q_2) < \alpha < 2(Q_1 + Q_2) - 2Q_1\ Q_2. \quad (10)$$

Оценить форму асимметрии возможно на основе выборочного коэффициента эксцесса, который сравнивает «крутость» выборочного распределения с нормальным распределением. Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю. Поэтому за стандартное значение выборочного коэффициента эксцесса принимают $E_k^* = 0$. Если $E_k^* < 0$, то кривая фактического распределения имеет более пологую вершину по сравнению с нормальной кривой; если $E_k^* > 0$, то вершина более крутая по сравнению с кривой нормального распределения.

Следует отметить, что вопросы качества данных в хранилищах неразрывно связаны с понятием *ETL*-процессов. *ETL*-процесс (от англ. extract, transform, load) – комплекс методов, реализующих извлечение данных из различных источников, их очистку, трансформацию и помещение в хранилище данных.

Вышеописанный механизм профилирования с точки зрения практического использования имеет смысл реализовать в виде этапа *ETL*-процесса, отвечающего за загрузку показаний приборов учета в хранилище биллинговой системы. В качестве основы для решения может быть выбрано любое из доступных *ETL*-средств, совместимых с технологиями, используемыми в конкретном хранилище.

Для очистки данных, основанных на технологиях СУБД SQL Server с успехом используется собственное (т.е. включенное в СУБД) решение Integration Services (SSIS), доступное, однако, не во всех редакциях продукта. В качестве альтернативы может быть применено программное средство с открытым исходным кодом, позволяющее решать *ETL*-задачи на основе функций библиотеки rhino-etl языка C#. Эта библиотека представляет процесс в виде конвейера, на каждом этапе которого применяется логика, нужным образом трансформирующая обрабатываемые строки.

В качестве решения для вычисления статистических показателей могут быть использованы возможности библиотеки с открытым кодом mathnet-numerics, которая позволяет легко определять требуемые параметры распределения с помощью специальных функций.

Таким образом, можно сделать следующие ***выводы***.

Вопросы качества данных в сфере расчетов за потребление энергоресурсов являются актуальными и достаточно важными. Часть проблем может быть решена путем совершенствования бизнес-логики в биллинговых системах. Остальные проблемы требуют применения ручного труда, который может быть частично автоматизирован с использованием методов анализа данных.

Возможности использования статистических метаданных для выявления аномальных данных о потреблении энергоресурсов в разрезе индивидуальных приборов учета и расчетных периодов на основе программных библиотек реализации ETL-процессов позволяют упростить решение задач интеграции биллинговых систем и АСКУЭ.

Список литературы

1. Айвазян С. А. Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – Москва : Финансы и статистика, 1982. – 465 с.
2. Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – Москва : Наука, 1965. – 412 с.
3. Зайнутдинова Л. Х. Управление энергосбережением бюджетных образовательных учреждений / Л. Х. Зайнутдинова // Прикаспийский журнал: управление и высокие технологии. – 2012. – № 2. – С. 164–170.
4. Любицын В. Н. Повышение качества данных в контексте современных аналитических технологий / В. Н. Любицын // Вестник Южно-Уральского государственного университета: компьютерные технологии, управление, радиоэлектроника. – 2012. – № 16. – С. 83–86.
5. Селко Д. SQL для профессионалов. Программирование / Д. Селко. – Москва : Лори, 2009. – 442 с.
6. О предоставлении коммунальных услуг собственникам и пользователям помещений в многоквартирных домах и жилых домов : постановление Правительства РФ от 06.05.2011 № 354 // Российская газета. – 2011. – Федеральный выпуск № 5492.
7. Шахгельдян К. И. Проблемы качества данных и информации в корпоративной информационной среде вуза / К. И. Шахгельдян. – Режим доступа: http://kis.vvsu.ru/userfiles/file/oiskp/quality.pdf (дата обращения 15.04.2014), свободный. – Заглавие с экрана. – Яз. рус.
8. Erhard Rahm. Data Cleaning: Problems and Current Approaches / Erhard Rahm, Hong Hai Do. – Available at: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf (accessed 15.04.2014).
9. Kriegel Hans-Peter. Outlier Detection Techniques / Kriegel Hans-Peter, Kroger Peer, Zimek Arthur. – Available at: http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf (accessed 15.04.2014).
10. Ng Andrew. Machine Learning and AI via Brain simulations / Ng Andrew. – Available at: http://www.ipam.ucla.edu/publications/gss2012/gss2012_10595.pdf/ (accessed 15.04.2014).

References

1. Ayvazyan S. A., Yenyukov I. S., Meshalkin L. D. *Prikladnaya statistika. Osnovy modelirovaniya i pervichnaya obrabotka dannykh* [Applied statistics. Modeling basics and initial data processing]. Moscow, Finansy i statistika, 1982. 465 p.
2. Bolshev L. N., Smirnov N. V. *Tablitsy matematicheskoy statistiki* [Tables of mathematical statistics]. Moscow, Nauka, 1965. 412 p.
3. Zaynutdinova L. Kh. Upravlenie energosberezheniem byudzhetnykh obrazovatelnykh uchrezhdeniy [Power management of budget educational institutions]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2012, no. 2, pp. 164–170.
4. Lyubitsyn V. N. Povyshenie kachestva dannykh v kontekste sovremennykh analiticheskikh tekhnologiy [Improvement in data quality in the context of modern analytical technologies]. *Vestnik Yuzhno-uralskogo gosudarstvennogo universiteta: kompyuternye tekhnologii, upravlenie, radioelektronika* [Bulletin of South Ural State University: computer technologies, management, radioelectronics], 2012, no. 16, pp. 83–86.
5. Selko D. *SQL dlja professionalov. Programmirovanie* [SQL for smarties. Programming]. Moscow, Lori, 2009. 442 p.
6. On the provision of utility services to owners and users of lodgements in apartment buildings and houses: The Resolution of the RF Government of 6 May, 2011, no. 354. Russian Newspaper, 2011, federal issue no. 5492. (In Russ.).
7. Shakhgeldyan K. I. Problemy kachestva dannykh i informatsii v korporativnoy informatsionnoy srede vuza [Data quality problems in university enterprise environment]. Available at: http://kis.vvsu.ru/userfiles/file/oiskp/quality.pdf (accessed 15 April 2014).
8. Erhard Rahm, Hong Hai Do. Data Cleaning: Problems and Current Approaches. Available at: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf (accessed 15 April 2014).
9. Kriegel Hans-Peter, Kroger Peer, Zimek Arthur. Outlier Detection Techniques. Available at: http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf (accessed 15 April 2014).
10. Ng Andrew. Machine Learning and AI via Brain simulations. Available at: http://www.ipam.ucla.edu/publications/gss2012/gss2012_10595.pdf/ (accessed 15 April 2014).