УДК 004.9

Р.А. Ешенко,

канд. техн. наук, доцент

Дальневосточного государственного университета путей сообщения

(г. Хабаровск)

Г.А. Гурвиц,

канд. техн. наук, доцент

Дальневосточного государственного университета путей сообщения

(г. Хабаровск)

ОЦЕНКА КАЧЕСТВА ДАННЫХ ПО ПРОИСХОЖДЕНИЮ, ПРОФАЙЛИНГ ДАННЫХ

Рассматривается понятие профайлинга как средства для оценки качества данных, разбираются основные ошибки, которые могут появляться в данных.

Ключевые слова: *профайлинг, качество данных, аномалии в данных, «грязные» данные.*

The term of «profiling» as a tool to evaluate the data quality is examined in the article. The basic errors that may occur in the data are analyzed.

Keywords: *profiling, data quality, anomalies in data, "dirty" data.*

В наше время профайлинг используется почти во всех сферах жизнедеятельности – от пассажирских перевозок до задач в бизнесе. Профайлинг позволяет выявлять аномалии в массиве данных и предотвращать повторения этих аномалий.

Профайлинг – это самое распространённое средство для оценки качества данных. В процессе производится множественная проверка полей на соответствие с заданными заранее параметрами и ограничениями. После выполнения данных проверок даётся оценка качеству данных. Если оценка неудовлетворительная, то необходимо произвести действия для улучшения и приведения к необходимому виду исходных данных. Профайлинг выполняется на основе анализа метаданных, которые описывают структуру данных.

Знание того, откуда берутся данные, является первым шагом в оценке их качества. Данные, в зависимости от того, кому они принадлежат, делятся:

1) на ваши собственные данные. Если вы рекламодатель, то это сведения, которые вы собираете о посетителях вебсайтов и клиентов CRM-систем. Если вы

издатель, то это данные, которые вы получаете непосредственно от посетителей вашего сайта. Эти данные обладают самым высоким качеством. Они лучше других данных хотя бы потому, что не будут стоить вам ни копейки;

2) данные рекламной активности, которые принадлежат кому-то, с кем вы взаимодействуете. Такие данные очень близки по своему качеству к вашим собственным данным, но их ценность зависит от уровня прозрачности;

3) сторонние данные, которые, как правило, имеют неизвестное происхождение. В редких случаях поставщик таких данных может указывать на то, откуда они «родом», но обычно это нельзя проверить.

В течение всего процесса профайлинга мы анализируем следующую информацию:

1) тип атрибута. Если тип не соответствует заданному параметру значения атрибута, мы должны принять меры, например в ячейке задано строковое значение, а ожидалось целочисленное;

2) длина значения. Если длина в значении атрибута превышает заданное число, надо применять меры для устранения данной ошибки;

3) дискретные значения. Обычно проверяется частота их появления и уникальность;

4) диапазон допустимых значений. Задаются ограничения значения, которые может принимать атрибут. При вводе некорректного значения можно предупреждать пользователя о несоответствии с ограничениями или при вводе таких значений менять его на значение «по умолчанию»;

5) анализ строковых шаблонов. Производим анализ ячейки строки с шаблоном, и, если совпадает, то ошибки нет, если не совпадает, необходимо включить обработку строки для её приведения к правилам.

При небольшом количестве данных для оценки можно применять визуальные методы. Для этого можно использовать как встроенные средства визуализации платформы, которую мы используем, так и сторонние программные решения.

Основными типами визуальной проверки являются оценка качества данных с помощью таблиц, оценка качества данных с помощью графиков.

Табличное представление помогает делать выводы о наличии нарушений, аномалий, пропусков и фиктивных значений. Все эти ошибки и несоответствия, очень выделяются на фоне окружающих данных. Числовые данные обычно выравниваются по правому краю ячейки таблицы, чтобы обеспечить расположение соответствующих разрядов чисел друг под другом (например, рубли под рублями, а копейки под копейками). Строковые данные, как правило, выравниваются по левому краю. Поэтому, если в столбце с числовыми данными окажется строковое значение, его можно легко обнаружить по выравниванию. Все эти ошибки можно заметить на табличном представлении

(рисунок 1).

Номер чека	Товар	Сумма
271512	КЕТЧУП, ХЛЕБ	58,70р.
272204	КОФЕ, САХАР, СЛИВКИ	0,25р.
2723	ЧАЙ, МОЛОКО	61,50р.
272350	КОЛБАСА, ХЛЕБ, МУКА	27580,45р
272351	ХЛЕБ, СЫР, МАСЛО	
2723512	МОРКОВЬ, КАПУСТА	45,16
272377	МОЛОКО	55,00р.

Некорректный номер чека

Аномальные значения

Пропущенное значение

Строковое значение

Рисунок 1 – Часто распространённые ошибки при заполнении таблиц

Если данных много, то такое решение сложно применить. Кроме того, сложно находить дубликаты и противоречия, так как они очень похожи на соседние ячейки и не выделяются в общей массе. Иногда данные проблемы можно решить, производя сортировку по определённым параметрам. Помимо этого, распространена оценка качества данных с помощью графиков. Графики и диаграммы позволяют выявить проблемы в данных, аномалии, шумы и пропущенные значения. При содержании шумов, график будет неравномерным. При содержании аномальных значений будут сильные отклонения в ту или другую сторону (рисунок 2).

Рисунок 2 – Пример графика с аномалиями и пропусками

Наличие проблем в данных можно обнаружить с помощью гистограмм распределения числовых значений, например сумм продаж. Как и многие реальные процессы, суммы покупок в супермаркетах распределены по нормальному закону, то есть существует диапазон значений, который встречается наиболее часто, в то время как вероятность появления очень малых сумм (менее 10 руб.) и больших (скажем, больше 5000 руб.) достаточно мала. Тогда кривая распределения сумм будет в той или иной степени соответствовать нормальному распределению. Появление в гистограмме дополнительных пиков (мод) может свидетельствовать о присутствии нехарактерных для данного ряда значений, которые мо-

гут являться аномалиями, ошибками и т.д. (рисунок 3).

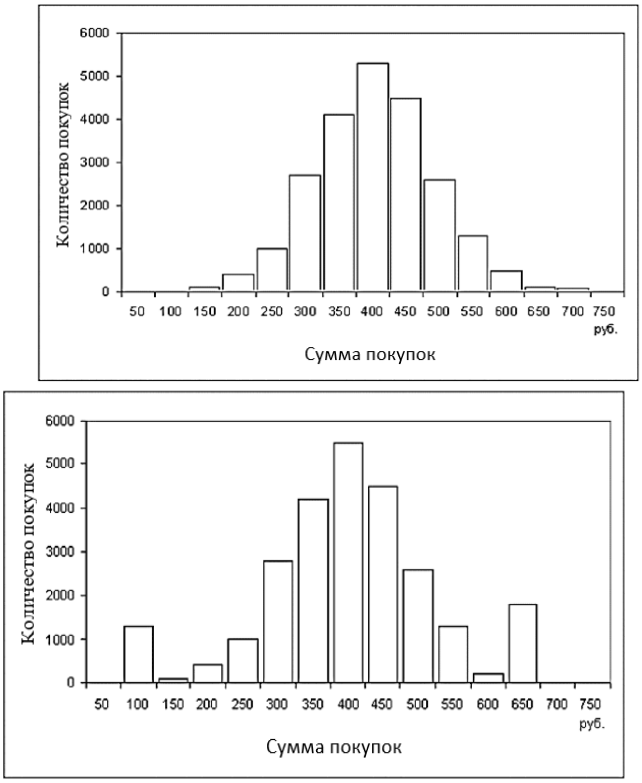

Рисунок 3 – Нормальное и аномальное распределение цен

Существует множество факторов, снижающих качество данных и не поддающихся формальной оценке. Один из главных факторов снижения качества данных – человеческий. Сотни и тысячи работников непрерывно формируют потоки «грязных» данных. При этом элементарные ошибки ввода и регистрации данных способны привести к большим искажениям информации о ходе бизнеса.

Главная задача оценки качества данных заключается в том, чтобы отличить «грязные» данные от чистых, то есть выявить, какие данные содержат ошибки, и оценить, какую долю «грязные» данные составляют в общем объёме данных. В ряде случаев эту задачу удаётся решить типовыми методами, например с помощью профайлинга. Однако для глубокого анализа причин загрязнения данных, как правило, требуются более изощрённые методы, основанные на знаниях о том, какими должны быть качественные данные. Такие знания содержат правила и шаблоны, которым должны отвечать качественные данные. Чем больше данных не соответствует шаблонам и правилам, тем сильнее они загрязнены и тем больше внимания требуется уделять очистке данных. Инструментальные средства анализа качества данных содержат десятки тысяч шаблонов и правил для выявления грязных данных.

Выделяются следующие основные классы ошибок:

– ошибки ввода, возникают в процессе ввода;

– ошибки в программном обеспечении;

– ошибки в базе данных системы учёта операций;

– ошибки консолидации данных, которые возникают в процессе переноса данных из филиалов в централизованное хранилище данных торговой организации. Эти ошибки связаны с тем, что каждый филиал имеет собственную БД для ведения учёта торговых операций, и представление данных в этих базах не всегда унифицировано. Например, для разных БД могут отличаться коды товаров и их групп, один и тот же код может быть привязан к различным торговым позициям и т.д. Все эти несоответствия проявляются в процессе переноса данных в хранилище, порождая структурные нарушения, противоречия, дубликаты и т.д.

Задача оценки качества данных заключается в выявлении ошибок и определении их критичности по отношению к результатам анализа. Обнаружение ошибок производится с помощью специальных процедур, каждая из которых разрабатывается для выявления определённого рода ошибок. При этом необходимо выбрать приоритет ошибок и обеспечить соответствующий порядок работы процедур. Результатом работы такой системы будет отчёт об обнаруженных ошибках и оценке уровня загрязнённости данных. Отчёт может использоваться аналитиками для разработки комплекса мер, направленных на повышение качества данных.

Оценка качества данных, контроль ошибок и анализ обнаруженных проблем – необходимые звенья любого проекта по бизнес-аналитике. Оценка качества данных не только позволяет определить степень их пригодности к анализу, но и служит индикатором оптимальности и продуманности работы систем.

Список использованных источников

1 Дюк В. А. Date Mining – интеллектуальный анализ данных / В. А. Дюк // Date Mining - Data Mining. - SPB.: Institute for Informatics and Automation of RAS Zhvirblis. Facebook keeps users "under the hood". URL: http://radio.bfm.ru (дата обращения 15.09.2017).

2 Microsoft SQL Server. Основные понятия интеллектуального анализа данных // URL: http://technet.microsoft.com (дата обращения 16.09.2017).

3 Николаи Д. Компания «StarPoint Software» придаст рекламе целевой характер / Д. Николаи // URL:https://www.osp.ru (дата обращения 16.09.2017).

4 Цветкова Р. Профайлер (Верификатор). ПрофГид / Р. Цветкова URL:https://www.profguide.ru (дата обращения 15.09.2017).

5 Автоматизация IP-сети. Ч. 3. Мониторинг TCP аномалий // URL:http://www.pvsm.ru (дата обращения 15.09.2017).