УДК 004.4

ПОВЫШЕНИЕ КАЧЕСТВА ДАННЫХ В КОНТЕКСТЕ СОВРЕМЕННЫХ АНАЛИТИЧЕСКИХ ТЕХНОЛОГИЙ

В.Н. Любицын

IMPROVEMENT IN DATA QUALITY IN THE CONTEXT OF MODERN ANALYTICAL TECHNOLOGIES

V.N. Lyubitsyn

Производится идентификация понятия «качество данных» применительно к информационно-аналитическим системам. Осуществляется деление методов повышения качества данных на группы и виды. Делается акцент на оценке качества данных как ключевом звене IT-технологий, связанных с управлением качеством данных. С целью формирования эффективной методики оценки данных предлагается классификация этапов процесса оценки качества данных и проводится их краткий сравнительный анализ, а также систематизация проблем качества данных на основе трех уровней с указанием для каждого из них специфики, включая соответствующие факторы, проявления и места борьбы с выявленными проблемами.

Ключевые слова: информационно-аналитическая система, хранилище данных, качество данных, ETL-процесс, очистка данных, предобработка данных, обогащение данных, методика оценки качества данных.

Identification of the term "data quality" with respect to the information-analytical systems is made in the article. Division of methods for enhancing the quality of the data into groups and species is carried out. Assessment of data quality as a key element of IT-related technology connected with management of data quality is emphasized. To form an effective methodology for assessing the data classification of stages of the process for assessing data quality is given, and brief comparative analysis is provided, data quality problems based on three levels, indicating the specificity of each of them, including the relevant factors, manifestations and control sites with identified problems, are systematized.

Keywords: information-analytical system, data warehouse, data quality, ETL-process, data cleaning, data preprocessing, data enrichment, data quality assessment tool.

Введение

Обеспечение требуемого качества данных, используемых в информационно-аналитической системе (ИАС) любого вида и назначения, почти всегда является одной из ключевых проблем создания подобной системы. В подтверждение этого тезиса проведем несложную аналогию, сравнив функцию качества информации с функцией фар автомобиля при движении ночью. Фары освещают дорогу впереди, обеспечивая контроль за обстановкой и подготовку реакции водителя, и если они светят слабо (информация неполная) или вбок (информация недостоверная), то неприятностей не избежать. В этой связи недостаточное внимание или уровень профессионализма при решении указанной проблемы может свести на нет все преимущества самых передовых и мощных методов и средств анализа, все усилия аналитиков и экспертов при подготовке управленческих решений из-за искажения истинной картины исследуемых бизнес-процессов, выявления ложных закономерностей, тенденций и связей между объектами бизнеса. Следствием этого станет выработка неверных управленческих решений, которые могут не только нанести ущерб, но и поставить под вопрос осуществление определенного вида деятельности и даже само существование организации, попавшей в такую ситуацию.

Следует отметить, что термин «качество данных» – «information quality», появившийся, кстати сказать, задолго до IT-технологий, в настоящее

Любицын Владимир Николаевич – кандидат технических наук, доцент кафедры информационно-аналитического обеспечения управления в социальных и экономических системах, Южно-Уральский государственный университет; lvn_iaou@mail.ru

Lyubitsyn Vladimir Nickolaevich – Candidate of Science (Engineering), Associate Professor of Information and Analytical Support in Social and Economic Systems Management Department, South Ural State University; lvn_iaou@mail.ru

время может трактоваться довольно широко и поразному в зависимости, в частности, от того, в какой области указанных технологий он употребляется. Применительно к ИАС, когда вопрос касается исходных данных, их качество следует понимать как совокупность свойств и характеристик этих данных, определяющих степень пригодности для последующего анализа [1].

Все многочисленные методы повышения качества данных целесообразно разделять на три группы, получившие в ряде источников [1–4] следующие названия:

– *очистка данных* – процесс выявления и исправления ошибок в исходной информации, т. е. оценка достоверности данных, выявление ошибочных подозрительных данных: аномалий, дубликатов, противоречий и т. п.;

– *предобработка данных* – процесс подготовки данных к решению конкретной аналитической задачи и приведение их в соответствие с требованиями, определяемыми спецификой этой задачи и способами ее решения, т. е. понижение размерности исходной информации, устранение незначащих признаков и т. п.;

– *обогащение данных* – процесс насыщения данных новой информацией, позволяющей сделать их более ценной для определенной аналитической задачи, т. е. привлечение информации из дополнительных источников, заполнение пропусков в информации, выявление связей между объектами и т. п.

При этом если методы очистки и предобработки данных можно целиком отнести к одному из этапов так называемого ETL-процесса (extraction, transformation, loading – извлечение, преобразование, загрузка), то с методами обогащения данных такой однозначности нет. Действительно, например, выявление связей между объектами связано с обработкой данных уже загруженных в хранилище данных (ХД) и предусматривает получение полезной информации, которая отсутствует в явном виде, но может быть получена с помощью манипуляций с имеющимися данными. Затем эта информация встраивается в виде новых полей или даже таблиц в ХД и может использоваться для дальнейшего анализа.

В этой связи представляется обоснованным разделять обогащение данных на два вида – внешнее и внутреннее. Внешнее обогащение данных, как правило, связано с решением стратегических бизнес-задач, требующих повышенного уровня аналитической работы. Именно в этом случае крайне необходимо в распоряжение аналитиков организации привлекать дополнительную информацию из внешних источников с тем, чтобы обогатить внутренние данные до уровня информативности и значимости, который позволит с высоким качеством решать задачи стратегического анализа. К внешним источникам данных следует отнести: другие организации, работающие в этой же сфере деятельности, причем как партнеры, так и конкуренты; органы государственной власти и местного самоуправления, включая налоговые и статистические службы; финансово-кредитные учреждения, банки, страховые компании; службы социальной сферы, включая органы труда и занятости, систему здравоохранения, пенсионный фонд.

Внутреннее обогащение данных не требует привлечения внешней информации, поскольку повышение информативности и значимости данных достигается за счет изменения их организации. Примером могут служить вычисленные и загруженные в ХД рейтинги сотрудников организации или оценки популярности товаров и т. д.

Важно понимать, что применение любого метода и, тем более, комплекса методов повышения качества данных, к какой бы группе или виду они не относились, требует предварительной оценки качества данных с целью выявления наиболее характерных проблем и уровня их сложности, а также выработки соответствующей стратегии по их решению. Здесь вполне уместен известный лозунг: «Предотвратить легче, чем исправить!» Ведь всегда проще и дешевле изначально застраховаться от проблем, чем потом лихорадочно исправлять ситуацию, теряя время, конкурентные преимущества, клиентов и, в конечном счете, доходы. Не случайно, что в IT-технологиях появилась новая дисциплина – управление качеством данных на предприятии (Enterprise Data Quality Management, EDQM). Более того, EDQM стало частью общего процесса управления качеством на предприятиях [5].

Ключевым звеном EDQM является именно оценка качества данных, которая реализуется на основе единовременной оценки, мониторинга или визуальной оценки. В любом случае разработка методики оценки качества данных требует ответа на вопрос: где именно ее следует проводить? При этом следует рассматривать следующие варианты: непосредственно в источниках данных, в ETL-процессе и в аналитической системе.

Первый из этих вариантов, т. е. оценка качества данных непосредственно в источниках данных, позволяет эффективно выполнить поиск орфографических ошибок, пропущенных, аномальных, логически неверных и фиктивных значений, противоречий и дубликатов на уровне записей и таблиц. Преимущества данного варианта в том, что результаты оценки качества данных, определенные методы очистки данных могут быть задействованы уже в ETL-процессе и в ХД поступят очищенные данные. Но надо помнить, что в ходе ETL-процесса качество данных может вновь ухудшится, поскольку происходит интегрирование данных из нескольких источников и могут появиться новые дубликаты и противоречия, несоответствия форматов и т. д. Следовательно, записи, уникальные и непротиворечивые для одного источника, могут потерять уникальность и непротиворечивость после объединения или слияния источников.

Второй вариант, т. е. оценка качества данных в ETL-процессе, в соответствии с выявленными проблемами и результатами оценки качества данных позволяет оперативно задействовать методы их очистки, загружая в ХД уже достоверную информацию. Хотя при этом возникает другая проблема, обусловленная тем, что использование данного подхода может заметно увеличить так называемое загрузочное окно, в течение которого существенно возрастает нагрузка на информационную систему организации.

В третьем случае, т. е. оценки качества данных в аналитической системе, а именно в процессе предобработки данных перед применением к ним различных методов Data Mining, эта оценка производится аналитиком визуально с использованием таблиц, графиков и диаграмм, а также на основе статистических оценок и характеристик. Действительно, например, с помощью гистограмм легко можно выявить аномальные значения, а оценка дисперсии позволяет оценить степень неравномерности ряда значений.

Безусловно, наиболее эффективным решением является использование всех трех вариантов, однако факторы времени и трудозатрат далеко не всегда позволяют «не мудрствуя лукаво» выбирать именно это. В любом случае не стоит забывать, что цель оценки качества данных – это лишь выявление в них каких-либо проблем (как правило, многочисленных), а локализация источников этих проблем и, тем более, борьба с ними должна осуществляться на других этапах повышения качества данных.

Другим важным аспектом формирования методики оценки качества данных следует считать необходимость классификации проблем, связанных с качеством данных, по отношению к одному из трех уровней: концептуальному, аналитическому или техническому. При этом наиболее критичными надо считать проблемы, отнесенные к концептуальному уровню. Ведь наличие подобных проблем свидетельствует о том, что стратегия сбора данных имеет серьезные пороки, а собранные и консолидированные данные в недостаточной мере отражают исследуемые бизнес-процессы. Если обнаружено, например, что данных недостаточно для всестороннего описания предметной области, то для решения проблемы необходимо использовать методы обогащения данных. Много реже оказывается, что объем данных избыточен, т. е. часть их иррелевантна по отношению к исследуемой предметной области, и нужно принимать меры по сокращению размерности исходного множества данных, уменьшая количество признаков и/или число их значений.

Такие факторы, как шумы данных, аномальные значения, противоречивые и дублирующие записи и пропуски, обусловливают проблемы качества данных, которые относят к аналитическому уровню. Однако следует учитывать, что для него весьма характерна субъективность оценки качества данных. Так, шум обычно проявляется в виде быстрых изменений значений ряда данных (скажем, объемов ежедневных продаж товара определенного вида), мешающих выявить общие закономерности и тенденции. Но то, что даже для одного и того же аналитика в одной ситуации будет просто шумом, в другом случае может считаться ценной информацией.

С аномалиями тоже не все так просто, поскольку бывает довольно трудно определенно утверждать, являются ли они лишь ошибками операторов или отражают реальные события, исключение которых ведет к потере важной информации. Наконец, идентификацию дублирующих записей нужно проводит весьма тщательно, ведь вполне вероятно, что два клиента с одинаковыми наименованиями и с разными адресами – это, на самом деле, совсем разные фирмы, в чем можно убедиться, дополнительно сравнив их банковские реквизиты.

К техническому уровню принято относить проблемы, связанные с нарушениями в структуре данных, их целостностью и полнотой, некорректностью форматов и кодировкой и т. п., что мешает интегрированию данных, их загрузке в ХД и в аналитические системы. Подобные проблемы достаточно просто выявляются по формальным признакам и ликвидируются.

Рассмотренная классификация проблем качества данных важна и для того, чтобы определиться с местом борьбы с ними. Проблемы технического уровня решаются только в ходе ETL-процесса, местом борьбы с проблемами аналитического уровня могут быть источники данных, ETL-процессы и аналитические системы, а проблемы концептуального уровня потребуют доработки стратегии сбора данных и/или аналитических процессов. В любом случае требуется внимание к проблемам каждого уровня, ведь, например, если остались проблемы концептуального уровня, то анализ накопленных данных оказывается совершенно бессмысленным, даже если они абсолютно корректны. Наличие в данных технических проблем, какую бы ценную информацию эти данные не содержали, просто не позволит предоставить ее аналитику, поскольку такие данные невозможно загрузить ХД. Напротив, данные, некорректные с точки зрения анализа, дойдут до аналитика, но вряд ли обрадуют его, поскольку не могут обеспечить значимые и достоверные результаты при использовании даже самых развитых аналитических технологий.

Что же касается конкретных технологий оценки качества данных, то вполне естественное стремление разработчиков ИАС минимизировать трудозатраты при повышении качества данных делает актуальным широкое использование так называемого профайлинга данных, в процессе которого анализируется следующая информация: тип, длина, шаблон и диапазон допустимых значе-

ний каждого атрибута (поля). Однако если объем исходных данных не слишком большой или среди них можно заранее определить наиболее значимую информацию, то для оценки качества данных не следует пренебрегать визуальными методами, используя для этого как встроенные средства визуализации, так и дополнительные программные инструменты. Конечно, «камнем преткновения» становятся трудно формализуемые ошибки, выявляемые с помощью более изощренных методов. Эти методы обычно требуют четких знаний о том, какими должны быть качественные данные, что далеко не всегда можно определить заранее. Именно в подобных случаях, когда нет каких-либо типовых решений, требуется не только профессионализм, но и творческий подход, поиск неординарных ходов по решению весьма нетривиальной задачи повышения качества данных. Наконец, никогда не стоит забывать и о таких простых, но достаточно эффективных способах борьбы за качество данных, как наличие четких, однозначно понимаемых технологических инструкций по вводу данных, поощрение сотрудников, допустивших наименьшее число ошибок, а также дублирование каналов ввода данных.

Заключение

Повышение качества данных – одна из наиболее важных и в то же время довольно сложных (в связи с трудностями формализации) задач ИАС, поскольку набор факторов, влияющих на качество данных, весьма разнообразен и может в процессе эксплуатации ИАС постоянно изменяться. Поэтому формированию и систематической модификации методики оценки качества используемых для анализа данных необходимо уделять большое внимание, поскольку именно она является основой для выбора места и технологии доведения качества данных до требуемого в конкретной ситуации уровня.

Литература

1. Ханк, Д.Э. Бизнес-прогнозирование: пер. с англ. / Д.Э. Ханк, Д.У. Уичерн, А.Д. Райтс. – 7-е изд. – М.: Издат. дом «Вильямс», 2003. – 651 с.

2. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян и др. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.

3. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б. Паклин. – СПб.: Питер, 2009. – 624 с.

4. Прикладная информатика: учеб. пособие / под ред. В.Н. Волковой и В.Н. Юрьева. – М.: Финансы и статистика: Инфра-М, 2008. – 768 с.

5. Ревякин, С.А. О важности качественной информации для принятия управленческих решений. – http://www.global-katalog.ru/cncat_jump.php?13146

Поступила в редакцию 2 апреля 2012 г.