Вычислительные технологии, 2024, том 29, № 1, с. 74–85. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 1, pp. 74–85. © FRC ICT, 2024 eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2024.29.1.007

Проблемы достоверности информации в OLAP-системах

А. А. Жирнов[1], И. Е. Харлампенков[1], О. Б. Кудряшова[2,*], В. П. Потапов[1]
[1]Федеральный исследовательский центр информационных и вычислительных технологий, 630090, Новосибирск, Россия
[2]Институт проблем химико-энергетических технологий СО РАН, 659322, Бийск, Россия
*Контактный автор: Кудряшова Ольга Борисовна, e-mail: olgakudr@inbox.ru
Поступила 28 октября 2022 г., доработана 02 декабря 2022 г., принята в печать 09 декабря 2022 г.

В настоящее время отсутствует единая система качественного и количественного определения таких атрибутов информации, как степень актуальности, достоверности и надежности источника данных. Целью работы является выработка подходов к оценке актуальности, достоверности информации и надежности источников данных в рамках информационной системы класса OLAP. Предлагаются единицы измерения, алгоритмы и методы вычисления достоверности информации и надежности источников. В дальнейшем выработанные подходы будут реализованы в виде алгоритмов и программ в рамках разрабатываемой информационной OLAP-системы.

Ключевые слова: актуальность информации, достоверность информации, надежность источников информации и данных, загрузчик данных, метаданные, OLAP-система, ETL/ELT-система.

Цитирование: Жирнов А.А., Харлампенков И.Е., Кудряшова О.Б., Потапов В.П. Проблемы достоверности информации в OLAP-системах. Вычислительные технологии. 2024; 29(1):74–85. DOI:10.25743/ICT.2024.29.1.007.

Введение

Популярная сегодня технология OLAP (online analytical processing) используется в автоматизированных системах предприятий для аналитической обработки данных, подготовки отчетов и разработки прогнозов на основе больших данных сложной структуры. Но в последние годы создаются информационные системы класса OLAP, не привязанные к системам управления конкретных предприятий, а предлагающих широкий комплекс услуг по предоставлению аналитических сводок и прогнозов, иногда с применением технологий машинного обучения и нейросетей. Так, например, развиваются гибридные системы, основанные на технологии OLAP, для поддержки регионального управления [1].

В системы OLAP входят подсистемы сбора и хранения данных, а также собственно подсистема анализа и прогнозирования. Между тем результаты аналитических расчетов напрямую зависят не только от объема, но от достоверности и актуальности исходных данных. Ценность знаний, полученных в результате интеллектуального анализа данных, зависит не только от используемых методов и алгоритмов анализа, но и от того, как подобраны и подготовлены исходные данные.

Отчасти достоверность и актуальность данных, предоставляемых слоем хранения OLAP-системы для анализа, обеспечиваются архитектурой этого слоя, которая предполагает проверку согласованности, непротиворечивости и хронологической целостности информации [2, 3]. С другой стороны, если мы говорим о фактографической и документальной информации, собираемой из открытых источников сети Интернет, важно оценивать ее достоверность и актуальность на этапе поиска и загрузки в базу данных.

В настоящей работе рассмотрены такие свойства информации, как актуальность, старение, точность, достоверность и надежность источника информации [4]. На схеме (рис. 1) главным параметром является время, так как от него зависят остальные свойства. Достоверность определяется тремя атрибутами: достаточностью (полнотой), целостностью и истинностью [5]. От надежности источника данных зависят остальные рассматриваемые свойства.

Большие современные OLAP-системы могут использовать одновременно десятки и сотни источников данных и представлять собой фабрики данных (обработка Big Data с применением искусственного интеллекта для построения, оптимизации алгоритмов и практического использования данных), озера данных (хранилища, содержащие обработанные и необработанные данные), экспертные системы с применением в них нейросетей.

В рассматриваемых системах используются данные (их часть) разных форматов и типов из различных источников информации. Эти данные динамически обновляются и относятся к категориям структурированных, полуструктурированных (слабоструктурированных) и неструктурированных данных, где они актуализируются (обновляются/добавляются) в режиме реального времени. Аналитическая подсистема не содержит средств ввода и редактирования данных, а работает с уже подготовленными консолидированными данными.

На этапе консолидации данных применяются системы ETL (извлечение, преобразование, загрузка) либо ELT (извлечение, загрузка, преобразование). Задача таких систем заключается не только в загрузке данных из различных источников, но и в их валидации, агрегировании и очистке. Под очисткой понимается устранение в данных аномалий, пропусков, дубликатов и противоречий. Для выявления аномалий (например, в данных временных рядов финансовой отчетности) используют статистические методы. Выявление дубликатов, пропусков и противоречий данных часто невозможно без

Рис. 1. Связь между точностью, актуальностью, достоверностью информации и надежностью источника данных

Fig. 1. A relationship between information accuracy, relevance and validity, and data source reliability

применения ручного труда. Самыми популярными системами типа ETL/ELT являются Airbyte, Azure, CloverETL, Apatar, Cloud Big Data, Apache Airflow и др.

В разных областях науки и техники понятия достоверности информации, надежности источников данных рассматриваются по-разному. Например, в работе [5] при оценке численных результатов решения задач разными методами предлагается в качестве численной характеристики достоверности использовать вероятность того, что фактическая погрешность не выходит за пределы, которые обусловлены вычислительной погрешностью. В некоторых случаях достаточно решения на основании выработанных критериев, в других — требуется более глубокий подход в количественном определении этих понятий. Например, определение более надежного источника данных из нескольких имеющихся [5].

В настоящей работе рассмотрены эти понятия применительно к информационным OLAP-системам. Развитие современных технологий приводит к углублению и уточнению в области обработки информации, получению более точных данных. Известны атрибуты, на основании которых можно судить о достоверности и актуальности информации: время создания; время изменения; источник данных; надежность источника данных, автор; кто изменил данные; сколько раз были изменены данные; на каком основании были внесены изменения и др. Главная проблема заключается в отсутствии системного подхода к оценке достоверности и актуальности информации или выработанных и систематизированных единиц измерения. Следует отметить, что информационная система определенного класса имеет свои требования к вышеперечисленным понятиям и должна проектироваться с учетом индивидуальных особенностей. Мы рассматриваем информационную систему класса OLAP, которая использует фактографические и документальные данные из широкого ряда внешних источников.

Целью работы является разработка подходов к оценке надежности источников данных, достоверности и актуальности информации, а также алгоритма для их вычисления для дальнейшей оценки достоверности отчетов и прогнозов. Этот алгоритм будет экспериментально апробирован в процессе разработки информационной OLAP-системы.

1. Описание данных информационной системы

В сети Интернет имеются как открытые, так и платные источники данных. Некоторые из них имеют свой внешний API, что значительно упрощает использование этих данных в своих целях. В разрабатываемой системе используются оба варианта этих источников данных. Данные содержат информацию: о компаниях в России с их финансовой отчетностью по годам; основных и неосновных видах деятельности компаний; таможенных данных; статистических данных по регионам России; курсах валют ЦБ РФ; месторождениях полезных ископаемых, их видах и типах; муниципальных образованиях и признаках; источниках этих данных; инвестиционных проектах; справочниках, таких как, например, Общероссийский классификатор организационно-правовых форм (ОКОПФ); Общероссийский классификатор видов экономической деятельности (ОКВЭД); Общероссийский классификатор форм собственности (ОКФС); Общероссийский классификатор органов государственной власти и управления (ОКОГУ); регионах России; товарной номенклатуре внешнеэкономической деятельности (ТНВЭД); странах мира; параметрах регионов России; единицах измерений и др. Однако некоторые данные содержатся в нескольких источниках, например данные о компаниях. В настоящее время данные о компаниях включают около 40 показателей. Возникает вопрос, какому из источников

больше верить. В каком источнике больше данных, отражающих достоверность, и по каким из этих атрибутов.

Форматы данных в перечисленных внешних источниках используются самые разнообразные (CSV, TXT, DOC, DOCX, DBF, PDF, JPEG, JSON, XML), а их типы охватывают почти все известные типы данных современных информационных систем от INTEGER до TEXT, в дальнейшем предполагается расширение числа источников информации. Таким образом, будет несколько источников данных со своими данными, которые не обязательно совпадают.

Процесс получения и обновления данных из разных источников сводится в первую очередь к достаточно большой степени автоматизации за счет использования ETL-системы Apache Airflow. Общий алгоритм ее работы сводится к настройке планировщика задач по времени, определению ссылок и местоположения данных, необходимой конвертации данных под требуемую структуру (удаление лишних полей, преобразование типов, арифметические действия с полями и т. д.) и самому процессу загрузки.

Процесс определения источников этих данных относительно детерминирован, однако для большей достоверности предполагается использовать несколько источников для сравнения информации и корректировки данных экспертами.

2. Подходы к оценке достоверности, актуальности информации и надежности источника данных

При использовании понятия достоверности информации следует обратиться к ее свойству отражать реально существующие объекты с необходимой точностью. Достоверной информацией считается та, которая не требует дополнительной проверки при ее использовании. Точность информации определяется степенью близости получаемой информации к реальному состоянию объекта, процесса, явления и т. п. Достоверность информации измеряется доверительной вероятностью необходимой точности, т. е. вероятностью того, что отображаемое информацией значение параметра отличается от его истинного значения в пределах необходимой точности [2].

В общем случае достоверность информации достигается путем:

- указания времени совершения события;
- исключением искаженной информации;
- своевременным вскрытием дезинформации;
- сопоставлением данных, полученных из различных источников.

Источник данных и его надежность напрямую влияют на остальные атрибуты информации, вследствие этого необходимо измерение этого свойства. Ненадежный источник информации порождает недостоверную информацию. Категории надежности источников информации определяются как: официальный, научный, надежный, проверенный, авторитетный и т. д. Даже в таких категориях имеется степень неопределенности, которая может меняться со временем.

Достоверность информации на основании определения описывается с помощью интеграла вероятностей Лапласа по формуле

$$P(|\tilde{x} - \overline{x}| \leq t\mu) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-t^2/2} dt,$$

где $t = (\tilde{x} - \overline{x})/\mu$ — нормированное отклонение выборочной средней от всей совокупности (коэффициент доверия); $\tilde{x}$ — значения выборочных средних; μ — среднее квадратическое отклонение; $\overline{x}$ — значение генеральной средней. Значение интеграла вероятностей (интеграла Лапласа) для разных t рассчитаны и приводятся в специальных таблицах [6].

В рамках требуемой задачи используются шкалы степени надежности источника и достоверности сведений, аналогичные предложенным в работе [7], однако с небольшими корректировками:

$$P(A)\begin{cases}\text{недостоверные данные} & \text{при } A = 0,\\ \text{низкая степень достоверности} & A = 0.25,\\ \text{достаточная степень достоверности} & A = 0.75,\\ \text{высокая степень достоверности} & A = 1.\end{cases}$$

Для того чтобы определить надежность источника, необходимо найти среднеарифметическое [6] ее составляющих [5] по формуле

$$\overline{x}_{\text{ap}} = \frac{1}{n}\sum_i x_i, \tag{1}$$

где n — число источников информации; x_i — надежность каждого источника в долях от единицы. В нашей работе не используется такая детальная оценка надежности, где каждому значению таблицы соответствует свой источник (1). Здесь каждой строке в таблице соответствует один источник.

Надежность источника информации означает качество информации, т. е. вероятность того, что информация о состоянии среды правильно идентифицирует среду как находящуюся в данном состоянии и определяется в диапазоне от 0 до 1 [8].

Формально шкала степени надежности источника с небольшими корректировками [7] для нашей задачи представлена в формуле

$$P(B)\begin{cases}\text{неизвестный источник} & \text{при } B \le 0.1,\\ \text{низкая степень надежности} & 0.1 < B \le 0.2,\\ \text{невысокая степень надежности} & 0.2 < B \le 0.3,\\ \text{удовлетворительная степень надежности} & 0.3 < B \le 0.4,\\ \text{средняя степень надежности} & 0.4 < B \le 0.5,\\ \text{достаточная степень надежности} & 0.5 < B \le 0.6,\\ \text{высокая степень надежности} & 0.6 < B \le 0.7,\\ \text{очень высокая степень надежности} & 0.7 < B \le 0.8,\\ \text{максимальная степень надежности} & 0.8 < B \le 0.9,\\ \text{абсолютная степень надежности} & 0.9 < B \le 1.\end{cases}$$

Итоговая достоверность вычисляется по формуле

$$P_{\text{total}} = P(A)P(B).$$

Актуальность определяется степенью сохранения ценности информации для управления в момент ее использования и зависит от динамики изменения ее характеристик и интервала времени, прошедшего с момента возникновения данной информации, т. е.

степенью соответствия информации текущему моменту времени, на основании которой вырабатываются шкалы степени соответствия. В разрабатываемой системе авторы используют две градации шкалы соответствия: актуальные данные и неактуальные.

Таким образом, для этих трех атрибутов информации представлены шкалы и количественные единицы измерения в рамках проектируемой системы.

3. Предлагаемые алгоритмы оценки достоверности, актуальности информации и надежности источника. Структура данных подсистемы

Разработана информационная подсистема оценки надежности источника данных (таблица метаданных — `source_reliability` — надежность источника данных; `data_sources` — источник данных), достоверности (таблица метаданных — `reliability` — достоверность) и актуальности данных (основная таблица исходных данных о компаниях `companies` и таблица `companies_history` для отслеживания изменений данных).

В связи с большим числом столбцов в некоторых таблицах, далее по тексту, в рисунках, будут использоваться значимые поля, влияющие на смысловую нагрузку. Остальные столбцы и строки (незначимые) будут обозначаться многоточием.

Реализация оценки надежности источника данных и достоверности предполагает использование математических критериев с дополнением человекопонятных, которые будут удобны пользователям информационных систем, а также двух таблиц: таблицы с источниками данных и таблицы метаданных. В таблице `source_reliability` используются поля: `id` — первичный ключ; `description` — человекопонятное поле для описания надежности; `range` — диапазон значений надежности (рис. 2).

В таблице `data_sources` имеются поля: `id` — первичный ключ; `title` — название источника данных; `reference` — URL-ссылка на веб-ресурс; `name` — название раздела источника информации; `description` — подробное описание источника информации; `year_id` — внешний ключ из другой таблицы; `source_reliability_id` — внешний ключ из таблицы `source_reliability`; `value_reliability` — само значение надежности источника информации (рис. 3).

Оценка достоверности информации предполагает использование математических критериев с дополнением человекопонятных, которые будут удобны пользователям информационной системы [9]. В таблице `reliability` используются поля: `id` — первичный ключ; `description` — человекопонятное описание достоверности; `range` — диапазон (рис. 4).

В таблице компаний содержится актуальная информация о компаниях: `id` — первичный ключ; `name` — название компании; `address` — адрес местонахождения компании и т. д.; `data_sources_id` — внешний ключ, источник информации; `data_sources_value` — внешний ключ, значение надежности источника информации; `reliability_id` — внешний ключ, значение достоверности; `final_reliability` — итоговая достоверность [10], вычисляемое поле (произведение полей `reliability_id` и `data_sources_value`) (рис. 5).

Реализация актуальности информации предполагается с использованием таблицы-дубликата (метод таблиц-двойников) `companies_history` и триггерных функций на удаление записей и обновление значений полей в таблице `companies` [11]. Считается, что свойство полноты информации одинаковое, так как все поля в исходной таблице

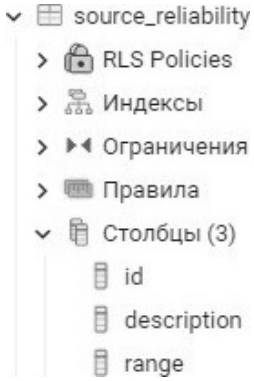

Рис. 2. Таблица метаданных для описания надежности источников данных
Fig. 2. Metadata table for description of data source reliability

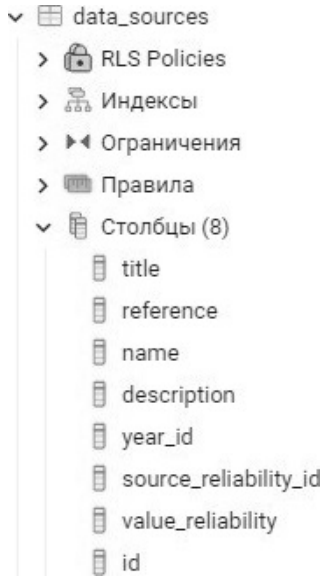

Рис. 3. Таблица описания источников информации и оценки их надежности при помощи внешнего ключа `source_reliability_id` и значения поля `value_reliability`
Fig. 3. Table for description of data sources and assessment of their reliability using the foreign key `source_reliability_id` and the value from the field `value_reliability`

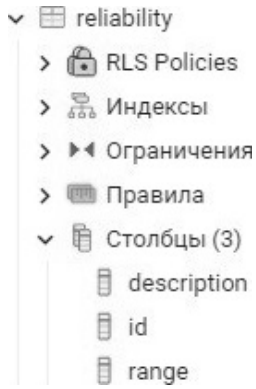

Рис. 4. Таблица метаданных для описания достоверности данных в таблице `companies`
Fig. 4. Metadata table for description of data validity in table `companies`

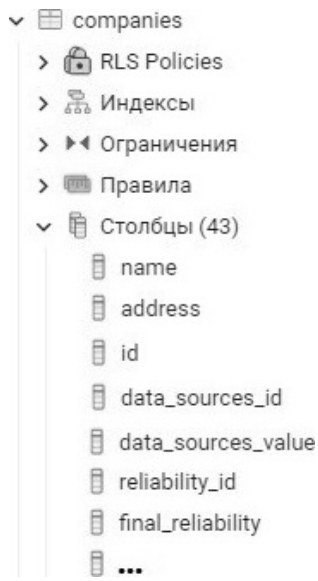

Рис. 5. Таблица данных о компаниях и оценка их достоверности
Fig. 5. Data table for companies, and estimates of data reliability

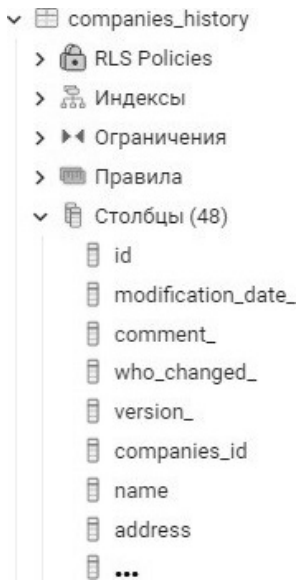

Рис. 6. Таблица неактуальных данных о компаниях
Fig. 6. Outdated data table for companies

(`companies`) полностью совпадают с полями таблицы изменений (`companies_history`) и к ним добавляются специальные поля (рис. 6).

На рис. 6 поля: `id` — первичный ключ; `modification_date_` — дата модификации значения в таблице `companies`; `comment_` — комментарий пользователя об изменении данных (на каком основании); `who_changed_` — внешний ключ того, кто изменил значение; `version_` — версия изменения информации (сколько раз было изменение); `companies_id` — внешний ключ из таблицы `companies`. Все остальные поля полностью дублируются из исходной таблицы `companies`, чтобы была возможность сквозного отслеживания изменений в ней всех значений.

Итоговая ER-диаграмма (рис. 7) отражает функционал отслеживания актуальности и достоверности данных, надежности источника информации и итоговой достоверности.

Пример реализации. В рамках выполнения требований установления надежности источника данных в подсистеме была создана градация, представленная в табл. 1.

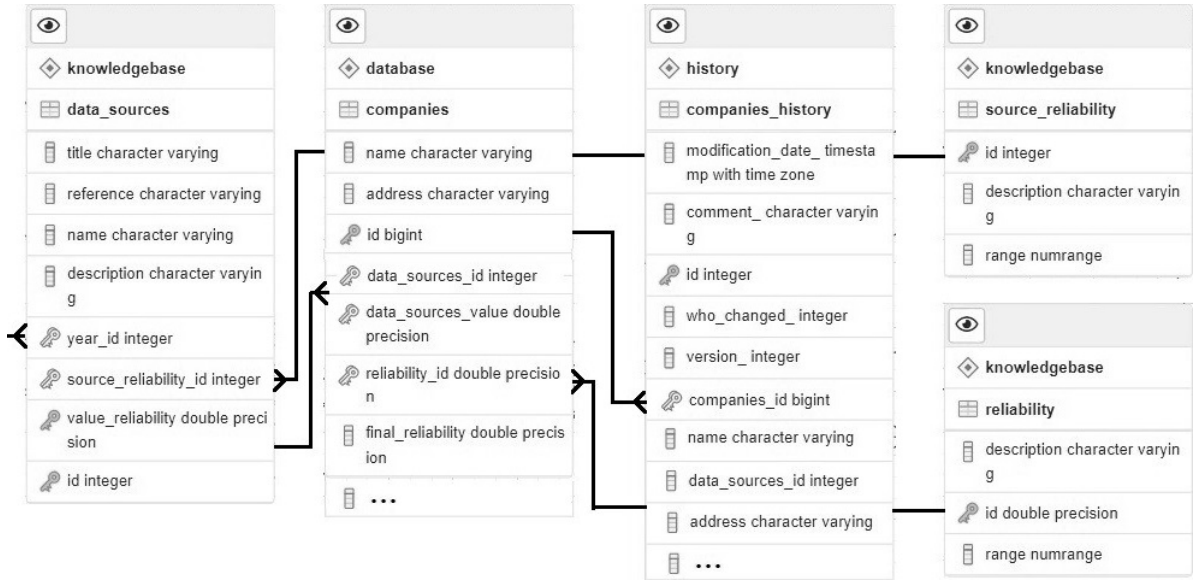

Рис. 7. Итоговая диаграмма сущность – связь
Fig. 7. An entity – relationship diagram

Т а б л и ц а 1. Установление значений надежности источника данных
Table 1. Measurement of data source reliability

Описание	Диапазон
Неизвестный источник	$[0; 0.1]$
Низкая степень надежности	$(0.1; 0.2]$
Невысокая степень надежности	$(0.2; 0.3]$
Удовлетворительная степень надежности	$(0.3; 0.4]$
Средняя степень надежности	$(0.4; 0.5]$
Достаточная степень надежности	$(0.5; 0.6]$
Высокая степень надежности	$(0.6; 0.7]$
Очень высокая степень надежности	$(0.7; 0.8]$
Максимальная степень надежности	$(0.8; 0.9]$
Абсолютная надежность	$(0.9; 1]$

Т а б л и ц а 2. Шкала значений достоверности данных
Table 2. Scale of data validity values

Описание	Диапазон
Недостоверные данные, ничтожная достоверность	$[0; 0.25]$
Низкая степень достоверности	$(0.25; 0.50]$
Достаточная степень достоверности	$(0.50; 0.75]$
Высокая степень достоверности	$(0.75; 1)$
Достоверные данные, наивысшая	$[1; 1]$

name character varying	address character varying	data_sources_id integer	data_sources_value double precision	reliability_id double precision	final_reliability double precision	...
ЭКСКЛЮЗИВН...	236023, Калин...	3	0.7	0.5	0.35	...
СК ХАЙНИКС Н...	129090, г. Моск...	2	0.8	0.5	0.4	...
СЕРВИС ТУРИЗ...	119002, г. Моск...	16	0.9	0.75	0.675	...
СТРОЙКЛЮЧ, О...	115191, г. Моск...	3	0.7	0.5	0.35	...
ТОРГОВЫЙ ДО...	105082, г. Моск...	22	0.7	0.25	0.175	...

Рис. 8. Таблица `companies`
Fig. 8. Table `companies`

modification_date_ timestamp with time zone	comment_ character varying	who_changed_ integer	version_ integer	companies_id_ bigint	name character varying
2022-06-12 08:01:08.1607...	comment	3	1	12860220	ТОРГОВЫЙ ДО...
2022-06-12 08:12:06.1660...	comment	3	1	12860340	СК ХАЙНИКС Н...
2022-06-12 08:30:36.5488...	comment	3	1	12860223	СЕРВИС ТУРИЗ...

Continuation of the table

address character varying	data_sources_id integer	reliability_id double precision	data_sources_value double precision	final_reliability double precision	...
105082, г. Моск...	3	0.5	0.7	0.35	...
129090, г. Моск...	3	0.5	0.7	0.35	...
119002, г. Моск...	3	0.5	0.7	0.35	...

Рис. 9. Таблица `companies_history`
Fig. 9. Table `companies_history`

В табл. 2 приведены критерии достоверности, которые требуются для пользователей разрабатываемой информационной системы.

О возможности отслеживания изменений во всех полях исходной таблицы свидетельствуют такие параметры информации, как актуальность/старение информации. На рис. 8 представлена исходная таблица с актуальной информацией компаний. Поля `modification_date_`, `comment_`, `who_changed_` и `version_`, добавленные к исходной таблице, указывают на актуальность информации (рис. 9). В табл. 1, 2 и рис. 8, 9 видна реализация и показаны данные актуальности, достоверности информации и надежности источника данных. Стоит отметить, что поля в таблицах имеют ссылочную целостность, поэтому оператор не сможет внести непреднамеренно неверные данные.

Таким образом, предложенный подход и алгоритм позволяют дать каждой строке таблиц с информацией численную оценку ее достоверности, актуальности, а также надежности источника данных. Как могут быть использованы указанные параметры на практике? Во-первых, при получении результата запроса пользователю может быть выдана информация о достоверности и актуальности предоставленных данных путем вычисления среднего арифметического численного значения этих полей в строках таблицы-результата запроса. Второй возможный вариант использования параметров заключается в их оценке при выполнении аналитических расчетов. В этом случае в алгоритм расчета следует включить значимость используемых данных (по шкале от 0 до 1). Итоговую достоверность результата расчета вычислить как взвешенное среднее с учетом значимости данных в качестве весов.

Заключение

В результате выполненной работы предложен подход к оценке достоверности, актуальности информации, а также надежности источника данных. Подход заключается в при-

менении необходимой шкалы градации этих свойств информации, а также количественной оценке достоверности и надежности источника данных. Приведен пример реализации подхода в разрабатываемой OLAP-системе через таблицу метаданных для достоверности информации, таблицу метаданных для оценки надежности источников данных и таблицу истории изменения полей для определения актуальности информации.

Алгоритм на основе таблиц-двойников и шкалы актуальности (определения изменений) позволяет отслеживать любое изменение значения ячейки, а не в целом записи в таблице базы данных, что является безусловным преимуществом предложенной реализации. В перспективе применение этих таблиц возможно для более глубокого анализа изменения данных с использованием нейронных сетей, аппроксимации, экстраполяции и прогнозирования.

Уровень достоверности и надежности источника информации определяется на основании всех имеющихся источников по данной тематике в информационной системе и экспертных оценок.

Результаты работы могут быть использованы в аналогичных системах и подходах к определению рассматриваемых свойств информации.

Благодарности. Работа выполнена в рамках государственного задания Минобрнауки России для Федерального исследовательского центра информационных и вычислительных технологий и Института проблем химико-энергетических технологий СО РАН.

Список литературы

[1] **Ноженкова Л.Ф.** Информационно-аналитические технологии и системы поддержки регионального управления. Вычислительные технологии. 2009; 6(14):71–81.

[2] **Орешков В.И. Паклин Н.Б.** Бизнес-аналитика: от данных к знаниям. Адрес доступа: `http://www.cfin.ru/itm/olap/cons.shtml`.

[3] **Melnykova N., Marikutsa U., Kryvenchuk U.** The new approaches of heterogeneous data consolidation. IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Sept. 11–14, 2018. Ukraine: Lviv; 2018: 408–411.

[4] **Еременко В.Т., Минаева В.А., Фисуна А.П., Коськина А.В.** Теория информации и информационных процессов: учебник для вузов. Орел: Госуниверситет — УНПК; 2015: 443.

[5] **Иванова С.М.** Оценка достоверности информации, найденной в сети Интернет. Преподаватель XXI век. 2015; 4(1):54–60.

[6] **Громыко Г.Л.** Теория статистики: учебник. 2-е изд., перераб. и доп. М.: ИНФРА-М; 2005: 476.

[7] **Лялькова Е.Е.** Информационные источники управленческого анализа. Управление экономическими системами. Электронный научный журнал. 2016; 8(90):25. Адрес доступа: `https://elibrary.ru/item.asp?id=26539597`.

[8] **Koops M.A.** Reliability and the value of information. Animal Behaviour. 2004; 1(67):103–111.

[9] **Жирнов А.А., Харлампенков И.Е., Потапов В.П.** Программа установления уровня достоверности источника данных. Адрес доступа: `https://new.fips.ru/ofpstorage/Doc/PrEVM/RUNWPR/000/002/022/618/984/2022618984-00001/document.pdf`.

[10] **Житников В.П., Шерыхалина Н.М.** Оценка достоверности численных результатов при наличии нескольких методов решения задачи. Вычислительные технологии. 1999; 6(4):77–87.

[11] **Жирнов А.А., Харлампенков И.Е., Юрченко А.В.** Программа сохранения изменений в ячейках одной таблицы базы данных в другую таблицу с записью версии. Адрес доступа: `https://new.fips.ru/ofpstorage/Doc/PrEVM/RUNWPR/000/002/021/666/299/2021666299-00001/document.pdf`.

Вычислительные технологии, 2024, том 29, № 1, с. 74–85. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 1, pp. 74–85. © FRC ICT, 2024 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2024.29.1.007

Information validity problems in OLAP systems

A. A. Zhirnov[1], I. E. Kharlampenkov[1], O. B. Kudryashova[2,*], V. P. Potapov[1]

[1]Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
[2]Institute for Problems of Chemical and Energetic Technologies SB RAS, 659322, Biysk, Russia
*Corresponding author: Olga B. Kudryashova, e-mail: `olgakudr@inbox.ru`
Received October 28, 2022, revised December 02, 2022, accepted December 09, 2022.

Abstract

The paper addresses the notion of information validity, relevance and data source reliability as applied to the design, development and information updating in contemporary OLAP systems. Many up-to-date information systems employ a variety of external data sources but the reliability of these sources and data validity therein raise doubts. In addition, the data is continually changed and updated from the old to the latest, which needs to be accounted. The problem is a lack of unified system for qualitative and quantitative assessment of information attributes such as degrees of relevance and validity combined with data source reliability. On the other hand, there is an array of information attributes that can judge relevance and validity of information. The present study aimed to elaborate approaches to assessing information relevance and validity and data source reliability for an OLAP database system and propose units of measurement, algorithms and computational methods. The elaborated approaches will be further employed as algorithms and programs as part of the developed OLAP database system.

Keywords: relevance of information, reliability of information, reliability of information and data sources, data loader, metadata, OLAP system, ETL/ELT system.

Citation: Zhirnov A.A., Kharlampenkov I.E., Kudryashova O.B., Potapov V.P. Information validity problems in OLAP systems. Computational Technologies. 2024; 29(1):74–85. DOI:10.25743/ICT.2024.29.1.007. (In Russ.)

Acknowledgements. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation for Federal Research Center for Information and Computational Technologies and for Institute for Problems of Chemical and Energetic Technologies of SB RAS.

References

1. **Nozhenkova L.F.** Information-analytical technologies and systems for support of regional management. Computational Technologies. 2009; 6(14):71–81. (In Russ.)
2. **Oreshkov V.I., Paklin N.B.** Biznes-analitika: ot dannykh k znaniyam [Business intelligence: from data to knowledge]. Available at: `http://www.cfin.ru/itm/olap/cons.shtml`. (In Russ.)

3. **Melnykova N., Marikutsa U., Kryvenchuk U.** The new approaches of heterogeneous data consolidation. IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Sept. 11–14, 2018. Ukraine: Lviv; 2018: 408–411.
4. **Eremenko V.T., Minaeva V.A., Fisuna A.P., Koskina A.V.** Teoriya informatsii i informatsionnykh protsessov: uchebnik dlya vuzov [Theory of information and information processes: a textbook for universities]. Orel: Gosuniversitet — UNPK; 2015: 443. (In Russ.)
5. **Ivanova S.M.** Assessment of the reliability of information found on the Internet. Teacher of the 21st Century. 2015; 4(1):54–60. (In Russ.)
6. **Gromyko G.L.** Teoriya statistiki [Theory of statistics: textbook. 2nd ed., revised]. Moscow: INFRA-M; 2005: 476. (In Russ.)
7. **Lyalkova E.E.** Information sources of management analysis. Management of Economic Systems: Electronic Scientific Journal. 2016; 8(90):25. Available at: `https://elibrary.ru/item.asp?id=26539597`. (In Russ.)
8. **Koops M.A.** Reliability and the value of information. Animal Behaviour. 2004; 1(67):103–111.
9. **Zhirnov A.A., Kharlampenkov I.E., Potapov V.P.** Programma ustanovleniya urovnya dostovernosti istochnika dannykh [A program for establishing the level of reliability of data source]. Available at: `https://new.fips.ru/ofpstorage/Doc/PrEVM/RUNWPR/000/002/022/618/984/2022618984-00001/document.pdf`. (In Russ.)
10. **Zhitnikov V.P., Sherykhalina N.M.** Certainty estimation of numerical results obtained by several methods of problem solution. Computational Technologies. 1999; 6(4):77–87. (In Russ.)
11. **Zhirnov A.A., Kharlampenkov I.E., Yurchenko A.V.** Programma sokhraneniya izmeneniy v yacheykakh odnoy tablitsy bazy dannykh v druguyu tablitsu s zapis'yu versii [A program for saving changes in cells of one database table to another table with a version record]. Available at: `https://new.fips.ru/ofpstorage/Doc/PrEVM/RUNWPR/000/002/021/666/299/2021666299-00001/document.pdf`. (In Russ.)