Очистка данных: проблемы и актуальные подходы

Эрхард Рам Хонг Хай До

Лейпцигский университет, Германия

http://dbs.uni-leipzig.de

Перевод: Осиповой Ю.Г.

Источник: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf">http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf

Мы классифицируем проблемы качества данных, которые рассматриваются в очистке данных и обеспечивают обзор основных подходов к решению. Очистка данных особенно необходимо при интеграции разнородных источников данных и должны решаться в комплексе со схемой связанных преобразованных данных. В хранилищах данных, очистка данных является важной частью так называемого ETL процесса. Мы также обсуждаем текущую поддержку инструментов для очистки данных.

Введение

Очистка данных имеет дело с выявлением и устранением ошибок и несоответствий в данных с целью улучшения качества данных. Проблемы качества данных присутствуют в отдельных наборах данных, таких как файлы и базы данных, например, из-за опечатки при вводе данных, недостачи информации или другие неверные данные. При интеграции множества источников данных, например, в хранилищах данных, интегрированных системах баз данных или глобальных информационных системах, необходимость в очистке данных существенно возрастает. Это потому, что источники часто содержат избыточные данные в различных представлениях. Для того чтобы обеспечить доступ к точным и последовательным данным, нужна консолидация различных представлений данных и устранение дублирования информации.

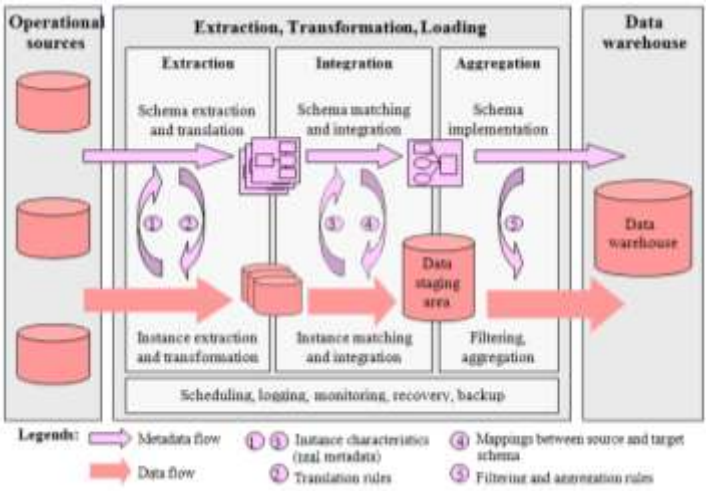

Рис 1 - Шаги построения Хранилища данных: ETL-процесс

Хранилища данных [6] [16] требуют и обеспечивают расширенную поддержку для очистки данных. Они загружают и постоянно обновляют большие объемы данных из различных источников, так что вероятность того, что некоторые источники содержат "грязные данные" является высокой. Кроме того, хранилища данных используется для принятия решений, таким образом, что правильность своих данных являются жизненно необходимым, чтобы избежать неправильных выводов. Например, дублирование или недостающая информация будет производить неправильную статистику ("мусор на входе, мусор на выходе").

В связи с широким спектром возможных несоответствий данных, а также большого объема данных, очистка данных, считается одной из самых больших проблем в хранилищах данных. Во время так называемого ETL процесса (извлечение, преобразование, загрузка), показана на рис. 1, дальнейшее преобразование данных связанно со схемой / преобразования данных и интеграцией, еще с фильтрацией и агрегацией данных, которые будут в Хранидищах данных. Как видно из рис. 1, вся очистка данных, как правило, выполняется в отдельной области подготовки данных до загрузки преобразованных данных в Хранилище. Большое количество инструментов различной функциональности для поддержки этих задач существуют на данный момент, но часто значительную часть работы по очистке и преобразованию данных нужно сделать вручную или с помощью низкоуровневых программ, которые трудны для написания и

поддерживания.

Интегрированные системы баз данных и информационные системы нуждаются в преобразовании данных, аналогичному, как и в хранилищах данных. В частности, там, как правило, для каждого источника данных для извлечения, и посредник для интеграции [32] [31]. До сих пор эти системы обеспечивали лишь ограниченную поддержку для очистки данных, вместо этого нужно сосредоточится на преобразования данных для перевода схемы и схемы интеграции. Данные не интегрируются заранее как для хранилищ данных, но необходимо извлечь из различных источников, преобразования и объединения в процессе отработки запросов. Соответствующие задержки передачи и обработки могут быть значительными, что затрудняет достижение приемлемого времени отклика. Усилия, необходимые для очистки данных при добыче и интеграция приведет к дальнейшему увеличению времени отклика, но являются обязательным для достижения полезных результатов запроса.

Подход очистка данных должна удовлетворять нескольким требованиям. Прежде всего, она должна обнаруживать и удалять все основные ошибки и несоответствия, как в отдельных источниках данных и при интеграции нескольких источников. Этот подход должен опираться на инструменты ограничиться ручной проверки и программирования усилий и быть расширяемой с легкостью покрыть дополнительные источники. Кроме того, очистка данных, не должны проводиться изолированно, а вместе со схемой данных, связанных с преобразованиями на основе всестороннего метаданных. Сопоставление функций для очистки данных и других преобразований данных должны быть определены декларативным образом и повторного использования для других источников данных, а также для обработки запросов. Специально для хранилищ данных, рабочих инфраструктуры должны поддерживаться, чтобы выполнить все шаги преобразования данных для нескольких источников и больших наборов данных в надежной и эффективной.

В то время как огромное количество исследований посвящена переводу схемы и схемы интеграции, очистки данных получил лишь немного внимания

научного сообщества. Ряд авторов сосредоточено на проблеме дублирующих выявление и устранение, например, [11] [12] [15] [19] [22] [23].Некоторые исследовательские группы сосредоточиться на общих проблемах, но не ограничиваясь, имеющие отношение к очистке данных, таких как специальные интеллектуального анализа данных подходов [30] [29], и преобразования данных на основе сопоставления схем [1], [21].Совсем недавно, несколько исследовательских усилий предложить и исследовать более полной и равномерной обработки данных очистки охватывает несколько этапов трансформации, конкретных операторов и их реализации [11] [19] [25].

В этой статье мы предлагаем обзор проблем, которые будут рассмотрены очистка данных и их решение. В следующем разделе мы приводим классификацию проблем. В разделе 3 обсуждаются основные подходы, используемые в очистке инструментов и научной литературы. В разделе 4 дается обзор коммерческий инструмент для очистки данных, в том числе ETL инструменты. Раздел 5 является заключение.

2 Данные проблемы очистки

Этот раздел классифицирует основные качества данных задач, решаемых данными очистки и преобразования данных. Как мы увидим, что эти проблемы тесно связаны и поэтому должны рассматриваться в едином порядке. Данные преобразования [26], необходимые для поддержки любых изменениях в структуре, представление или содержание данных. Эти преобразования возникла необходимость во многих ситуациях, например, иметь дело со схемой эволюции, миграции унаследованных систем на новую информационную систему, или при наличии нескольких источников данных должны быть интегрированы.

Как показано на рис. 2 грубо различать одного источника и нескольких источников проблем и между схемой и экземпляром проблем. Схема уровня проблемы, конечно, отражены также в случаях, они могут быть решены на уровне схемы по усовершенствованной конструкции схемы (схема эволюции), схему перевода и схемы интеграции. Экземпляр на уровне проблем, с другой

стороны, ссылаются на ошибки и несоответствия в фактическое содержание данных, которые не видны на уровне схемы. Они являются основной акцент очистки данных. Рис. 2 также указывает на некоторые типичные проблемы для различных случаев. Хотя это и не показано на рис. 2, из одного источника возникновения проблем (с повышенной вероятностью) в нескольких источников случае, также, помимо конкретных нескольких источников проблем.

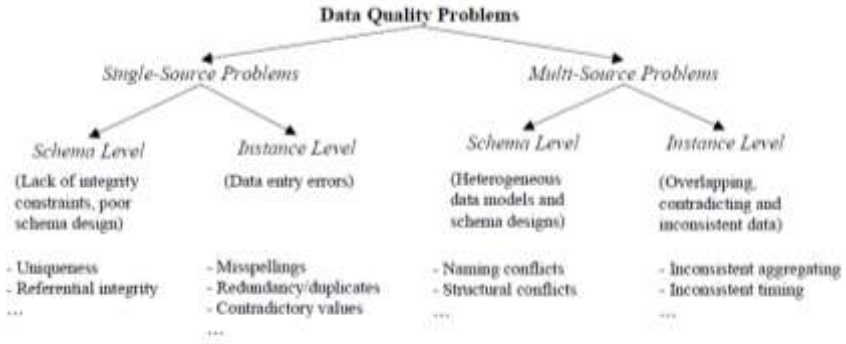

Figure 2. Classification of data quality problems in data sources

Рисунок 2. Классификация проблем качества данных в источниках данных

2.1 Примеры проблем отдельного источника данных на уровне схемы (нарушение ограничения целостности)

Качества данных источника в значительной степени зависит от того, в какой степени он определяется схемой и ограничений целостности управления допустимое значение данных. Для источников без схемы, такие как файлы, есть несколько ограничений на то, что данные могут быть введены и сохранены, что приводит к высокой вероятности ошибок и несоответствий. Системы управления базами данных, с другой стороны, соблюдение ограничений конкретной модели данных (например, реляционный подход требует простых значений атрибутов, ссылочной целостности и т.д.), а также конкретных приложений ограничений целостности. Схема проблем, связанных с качеством данных таким образом возникают из-за отсутствия соответствующих конкретной модели или конкретного приложения ограничений целостности, например, в связи с моделью данных ограничений или плохой дизайн схемы, или потому, что только несколько ограничений целостности, были

определены ограничить накладные расходы на контроль целостности. Экземпляр конкретные проблемы касаются ошибок и несоответствий, которые не могут быть предотвращены на уровне схемы (например, опечатки).

Область / Проблема		Грязные данные	Причины / Примечания
атрибут	Недопустимые значения	день рождение=30.13.70	Значения за пределами областиспектра
запись	Нарушается зав исимости атриб утов	возраст=22, лень рождения=12.02.70	возраст = (текущая дата - дата рождения) следует проводить
тип записи	Уникально сть наруше ние	emp_1=(имя = ”Смит Д.”, ИИН=”123456”) emp_2=(имя =”Петр Миллер”, SSN=”123456”)	единственности для ИИН (номер социального страхования) нарушены
источник	Нарушение ссылочной	emp=(имя=”Джон Смит”, отдел=127)	ссылка отдел (127) не определена

Таблица 1. Примеры из одного источника проблемы на уровне схемы (нарушение ограничения целостности)

Для обеих схем и на уровне экземпляра проблемы мы можем дифференцировать различные области проблема: атрибут (поле), запись, тип записи и источник, примеры для различных случаев приведены в таблицах 1.

2. Обратите внимание, что указанные ограничения уникальности на уровне схемы не мешает дублировать случаях, например, если информация на том же реального объекта вводится дважды с разными значениями атрибута. Учитывая, что очистка источников данных является дорогостоящим процессом, предотвращая грязных данных, который необходимо ввести, очевидно, важным шагом к снижению очистка проблемы. Это требует соответствующей конструкции базы данных и ограничений целостности, а также приложения для ввода данных. Кроме того, открытие очистки данных правил при проектировании склада может предложить улучшения ограничений в структуру существующих схем.

2.2 Multi-источник проблемы

Проблемы присутствуют в одном источников усугубляется, когда несколько источников должны быть интегрированы. Каждый источник может содержать грязные данные и данные в источниках могут быть представлены по-разному, дублирование или противоречие. Это потому, что источники, как правило, развиты, развертывания и

обслуживания независимо служить конкретным потребностям. Это приводит к большой степени неоднородности относительно системы управления данными, модели данных, схемы проектов и фактических данных.

На уровне схемы, модели данных и различных схем, которые будут рассмотрены по шагам по переводу схемы и схемы интеграции, соответственно. Основные проблемы w.r.t. Схема конструкции наименований и структурные конфликты [2], [24] [17]. Именование конфликты возникают, когда имя используется для различных объектов (омонимы) или различные имена используются для того же объекта (синонимы). Структурные конфликты возникают в различных вариантах и ссылаться на различные представления одного и того же объекта в разных источниках, например, атрибут против табличного представление, различные структуры компонентов, различные типы данных, различные ограничения целостности данных и т.д.

В дополнение к схеме, уровень конфликтов, многие конфликты возникают только на уровне экземпляра (по данным конфликтов). Все проблемы из одного источника могут произойти с различными представлениями в различных источниках (например, дублированных записей, противоречащие записи). Кроме того, даже тогда, когда одни и те же имена атрибутов и типов данных, могут быть различными представления значения (например, семейное положение), или иная интерпретация значения (например, единицы измерения доллара против евро) из разных источников. Кроме того, информация в источниках могут быть предоставлены на различных уровнях агрегации (например, продажи каждого продукта по сравнению с продажи каждого продукта группы) или обращение к различным моментам временим(например, текущие продажи по состоянию на вчера источник 1 - на прошлой неделе источник 2).

Главной проблемой для очистки данных из нескольких источников является выявление перекрывающихся данных, в частности, соответствующие записи, относящиеся к той же сущности реального мира (например, клиент). Эта проблема также называется проблемой идентичности объекта [11], устранения дублирования или слияния/ чистка проблемы [15]. Часто, информация лишь

частично избыточных и источников могут дополнять друг друга, предоставляя дополнительную информацию об объекте. Таким образом, дублирование информации должны быть очищены, и дополняет информацию должны быть консолидированы и объединены в целях обеспечения согласованного представления реальных объектов.

Покупатели (источник 1)

ПИН	Название	Улица	Город	Пол
11	Кристин Смит	2 Харлей улl	Южный Форк,48503	0
24	Кристиан Смит	Улица Харлейу	Ю. Форк	1

Клиенты (источник 2)

Cno	*Фамилия*	*Имя*	*Пол*	*Адресс*	*Телефон*
24	Смит	Кристин	М	23 ул. Харлей, Чикаго, 60633-	333-222-6542
493	Смит	Кристи	Ж	2 Харлей, Юг	444-555-6666

Покупатели (интегрированные поля с очисткой данных)

№	*Фамилия*	*Имя*	*Пол*	*Улица*	*Город*	Телефон	*Fax*	*ПИН*	*Cno*
1	Smith	Kristen L.	F	2 Hurley	South	444-555-		11	493
2	Smith	Christian	M	2 Hurley	South			24	
3	Smith	Christoph	M	23 Harley	Chicago	333-222-	333-222-		24

Таблица 2. Примеры из различных источников проблем на схеме и на уровне экземпляра

Два источника в примере на рис. 3 представлены в реляционном формате и отображают конфликты. На уровне схемы, есть конфликты имен (синонимы Покупатель/Клиент) и структурные конфликты (различные представления имен и адресов). На уровне экземпляра, отметим, что существуют различные представления пола (("0"/"1" против "F"/"M") и,вероятно, дублирующиеся записи (Кристен Смит).Последние н аблюдения также показывают, что в то время как *Cid*/*Cno* являются источником конкретных идентификаторов, их содержание не сопоставимы между источниками; разных номеров (11/493) могут относиться к одному человеку в то время как разные люди могут иметь такое же количество (24) . Решение этих проблем требует не только интеграции схем и очистки данных, третья таблица показывает возможное решение. Отметим, что схема конфликты должны быть урегулирована первой позволяют очистки данных, в частности, обнаружение дубликатов на

основе единого представления имен и адресов, а также согласование Пол /Sex значения.

Выводы

Мы предоставили классификации проблем качества данных в источниках данных между одним и несколькими источниками, а также между схемой и на уровне экземпляра проблем. Мы также очертили основные шаги по очистке данных и преобразование данных, и подчеркнули необходимость, чтобы покрыть схемы и экземпляр данных, связанных с преобразованиями на основе комплексного подхода. Кроме того, мы представили обзор коммерческих данных чистящие средства. Несмотря на то, о состоянии дел в этих средств достаточно продвинутые, они обычно покрывают только часть проблемы, и по-прежнему требует значительных усилий руководства или само-программирования. Кроме того, их взаимодействие ограничено (собственный API, и метаданные, представлений). До сих пор лишь небольшое исследование появилось на очистке данных, несмотря на большое количество инструментов указывает и на важность и сложность проблемы очистки. Мы видим несколько тем, заслуживающих дальнейшего исследования. Прежде всего, необходимо активизировать работу по разработке и внедрению лучших язык подход к поддержке и схему, и преобразования данных. Например, операторы, такие как Match, Merge или сопоставление композиции были либо учились в пример (данные) или схемы(метаданных) уровне, но могут быть построены на тех же методов реализации. Очистка данных нужна не только для хранения данных, но и для обработки запросов на гетерогенных источников данных, например, в веб-информационных систем. Эта среда создает гораздо более строгие ограничения производительности для очистки данных, которые необходимо учитывать в разработке приемлемых подходов. Кроме того, очистки данных для частично структурированных данных, например, основанные на XML,может иметь большое значение, учитывая снижение структурных ограничений и быстро увеличивающееся количество XML-данных.

Благодарности

Мы хотели бы поблагодарить Фил Бернштейн, Елена Galhardas и Сунита Sarawagi за полезные замечания.

Referenes

[1] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: Tools for Data Translation and Integration. In
[26]:3-8, 1999.

[2] Batini, C.; Lenzerini, M.; Navathe, S.B.: A Comparative Analysis of Methodologies for Database Schema Integra-
tion. In Computing Surveys 18(4):323-364, 1986.

[3] Bernstein, P.A.; Bergstraesser, T.: Metadata Support for Data Transformation Using Microsoft Repository. In
[26]:9-14, 1999

[4] Bernstein, P.A.; Dayal, U.: An Overview of Repository Technology. Proc. 20th VLDB, 1994.

[5] Bouzeghoub, M.; Fabret, F.; Galhardas, H.; Pereira, J; Simon, E.; Matulovic, M.: Data Warehouse Refreshment. In
[16]:47-67.

[6] Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record
26(1), 1997.

[7] Cohen, W.: Integration of Heterogeneous Databases without Common Domains Using Queries Based Textual
Similarity. Proc. ACM SIGMOD Conf. on Data Management, 1998.

[8] Do, H.H.; Rahm, E.: On Metadata Interoperability in Data Warehouses. Techn. Report, Dept. of Computer Sci- ence, Univ. of Leipzig. http://dol.uni-leipzig.de/pub/2000-13.

[9] Doan, A.H.; Domingos, P.; Levy, A.Y.: Learning Source Description for Data Integration. Proc. 3rd Intl. Work- shop The Web and Databases (WebDB), 2000.

[10] Fayyad, U.: Mining Database: Towards Algorithms for Knowledge Discovery. IEEE Techn. Bulletin Data Engi- neering 21(1), 1998.

[11] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: Declaratively cleaning your

data using AJAX. In Journees
Bases de Donnees, Oct. 2000. http://caravel.inria.fr/~galharda/BDA.ps.
[12] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: AJAX: An Extensible Data Cleaning Tool. Proc. ACM SIG- MOD Conf., p. 590, 2000.
[13] Haas, L.M.; Miller, R.J.; Niswonger, B.; Tork Roth, M.; Schwarz, P.M.; Wimmers, E.L.: Transforming Heteroge- neous Data with Database Middleware: Beyond Integration. In [26]:31-36, 1999.
[14] Hellerstein, J.M.; Stonebraker, M.; Caccia, R.: Independent, Open Enterprise Data Integration. In [26]:43-49,
1999.
[15] Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data
Mining and Knowledge Discovery 2(1):9-37, 1998.
[16] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Springer, 2000.
[17] Kashyap, V.; Sheth, A.P.: Semantic and Schematic Similarities between Database Objects: A Context-Based Ap- proach. VLDB Journal 5(4):276-304, 1996.
[18] Lakshmanan, L.; Sadri, F.; Subramanian, I.N.: SchemaSQL – A Language for Interoperability in Relational Multi- Database Systems. Proc. 26th VLDB, 1996.
[19] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: Cleansing Data for Mining and Warehousing. Proc. 10th Intl. Conf.
Database and Expert Systems Applications (DEXA), 1999.
[20] Li, W.S.; Clifton, S.: SEMINT: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases
Using Neural Networks. In Data and Knowledge Engineering 33(1):49-84, 2000.
[21] Milo, T.; Zohar, S.: Using Schema Matching to Simplify Heterogeneous Data Translation. Proc. 24th VLDB,
1998.
[22] Monge, A. E. Matching Algorithm within a Duplicate Detection System. IEEE Techn. Bulletin Data Engineering
23 (4), 2000 (this issue).
[23] Monge, A. E.; Elkan, P.C.: The Field Matching Problem: Algorithms and

Applications. Proc. 2nd Intl. Conf.
Knowledge Discovery and Data Mining (KDD), 1996.
[24] Parent, C.; Spaccapietra, S.: Issues and Approaches of Database Integration. Comm. ACM 41(5):166-178, 1998. [25] Raman, V.; Hellerstein, J.M.: Potter's Wheel: An Interactive Framework for Data Cleaning. Working Paper, 1999.
http://www.cs.berkeley.edu/~rshankar/papers/pwheel.pdf.
[26] Rundensteiner, E. (ed.): Special Issue on Data Transformation. IEEE Techn. Bull. Data Engineering 22(1), 1999. [27] Quass, D.: A Framework for Research in Data Cleaning. Unpublished Manuscript. Brigham Young Univ., 1999 [28] Sapia, C.; Höfling, G.; Müller, M.; Hausdorf, C.; Stoyan, H.; Grimmer, U.: On Supporting the Data Warehouse
Design by Data Mining Techniques. Proc. GI-Workshop Data Mining and Data Warehousing, 1999.
[29] Savasere, A.; Omiecinski, E.; Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Data- bases. Proc. 21st VLDB, 1995.
[30] Srikant, R.; Agrawal, R.: Mining Generalized Association Rules. Proc. 21st VLDB conf., 1995.
[31] Tork Roth, M.; Schwarz, P.M.: Don’t Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources. Proc. 23rd VLDB, 1997.