Тема 8. Основи аналізу даних

План

1. Підготовчі етапи процесу Data Mining.
2. Дублювання даних.
3. Очищення даних.
4. Етапи очищення даних.

Мета вивчення теми: вивчити основи інтелектуального аналізу даних; засвоїти, в чому полягає процес очищення даних.

Перелік ключових слів та понять із теми

Інтелектуальний аналіз даних, дослідження, предметна область, завдання, дані, набір даних, очищення даних, дублювання даних, викид

Теоретичні відомості з теми

1. Підготовчі етапи процесу Data Mining.

Процес Data Mining є свого роду дослідженням. Як будь-яке дослідження, цей процес складається з певних етапів, що включають елементи порівняння, типізації, класифікації, узагальнення, абстрагування, повторення.

Процес Data Mining нерозривно пов’язаний із процесом прийняття рішень.

Процес Data Mining будує модель, а в процесі прийняття рішень ця модель експлуатується.

Розглянемо традиційний процес Data Mining. Він включає такі етапи:

- аналіз предметної області;
- постановка задачі;
- підготовка даних;
- побудова моделей;
- перевірка й оцінка моделей;
- вибір моделі;
- застосування моделі;
- корекція й відновлення моделі.

Етап 1. Аналіз предметної області.

Дослідження – це процес пізнання певної предметної області, об’єкта або явища з певною метою.

Процес дослідження полягає в спостереженні властивостей об’єктів з метою виявлення й оцінки важливих, з погляду суб’єкта-дослідника, закономірних відносин між показниками даних властивостей.

Вирішення будь-якого завдання у сфері розробки програмного забезпечення повинне починатися з вивчення предметної області.

Предметна область – це подумки обмежена область реальної дійсності, що підлягає опису або моделюванню й дослідженню.

Предметна область складається з об’єктів, що різняться за властивостями й перебувають у певних відносинах між собою або взаємодіють яким-небудь чином. Дослідникові необхідно вміти виділити їхню частину, необхідну для використання. Наприклад, при розв’язанні задачі «Чи видавати кредит?» важливими є всі дані про приватне життя клієнта, аж до того, чи має роботу подружжя, чи є в клієнта неповнолітні діти, який його рівень освіти тощо. Для розв’язку іншого завдання банківської діяльності ці дані будуть абсолютно неважливі. Отже, важливість даних залежить від вибору предметної області.

У процесі вивчення предметної області повинна бути створена її модель. Знання з різних джерел повинні бути формалізовані за допомогою яких-небудь засобів.

Це можуть бути текстові описи предметної області або спеціалізовані графічні нотації. Існує велика кількість методик опису предметної області, *наприклад,* методика структурного аналізу SADT і заснована на ньому IDEF0, діаграми потоків даних Гейна-Сарсона, методика об’єктно-орієнтованого аналізу UML та ін. Модель предметної області описує процеси, що відбуваються в предметній області, і дані, які в цих процесах використовуються.

Це перший етап процесу Data Mining. Але від того, наскільки вірно змодельована предметна область, залежить успіх подальшої розробки додатка Data Mining.

Етап 2. Постановка задачі.

Постановка задачі Data Mining включає такі кроки:

- формулювання задачі;
- формалізація задачі.

Постановка задачі включає також опис статичної та динамічної поведінки досліджуваних об’єктів.

Приклад. При просуванні нового товару на ринок необхідно визначити, яка група клієнтів фірми буде найбільш зацікавлена в цьому товарі.

Опис статики – це опис об’єктів та їх властивостей.

Приклад. Клієнт є об’єктом. Властивості об’єкта «клієнт»: родинний стан, дохід за попередній рік, місце проживання.

При описі динаміки описується поведінка об’єктів і причини, що впливають на їхню поведінку.

Приклад. Клієнт купує товар А. З появою нового товару В клієнт уже не купує товар А, а купує тільки товар В. Поява товару В змінила поведінку клієнта. Динаміка поведінки об’єктів часто описується разом зі статикою.

Технологія Data Mining не може замінити аналітика й відповісти на питання, що не були задані. Тому постановка задачі є необхідним етапом процесу Data Mining, оскільки саме на цьому етапі визначається, яку ж задачу необхідно розв’язати. Іноді етапи аналізу предметної області й постановки задачі поєднують в один етап.

Етап 3. Підготовка даних.

Ціль етапу: розробка бази даних для Data Mining (поняття «даних» було розглянуто в темі 2).

Підготовка даних є найважливішим етапом, від якості виконання якого залежить можливість одержання якісних результатів усього процесу Data Mining. Крім того, слід пам'ятати, що на етап підготовки даних, за деякими оцінками, може бути витрачене до 80% усього часу, відведеного на проект.

Розглянемо докладно цей етап.

1. ***Визначення й аналіз вимог до даних.*** На цьому кроці здійснюється так зване моделювання даних, тобто визначення й аналіз вимог до даних, які необхідні для здійснення Data Mining. При цьому вивчаються питання розподілу користувачів (географічне, організаційне, функціональне); питання доступу до даних, які потрібні для аналізу, необхідність у зовнішніх або внутрішніх джерелах даних; а також аналітичні характеристики системи (виміру даних, основні види вихідних документів, послідовність перетворення інформації тощо).

2. ***Збір даних.*** Наявність в організації сховища даних робить аналіз простіше й ефективніше, його використання, з погляду вкладень, обходиться дешевше, ніж використання окремих баз даних або вітрин даних. Однак далеко не всі підприємства оснащені сховищами даних. У цьому випадку джерелом для вхідних даних є оперативні, довідкові й архівні БД, тобто дані з існуючих інформаційних систем.

Також для Data Mining може знадобитися інформація з інформаційних систем керівників, зовнішніх джерел, паперових носіїв, а також знання експертів або результати опитувань.

Слід пам'ятати, що в процесі підготовки даних аналітики й розроблювачі не повинні прив'язуватися до показників, які є в наявності, й описати максимальну кількість факторів і ознак, що впливають на процес, що аналізується.

На цьому етапі здійснюється кодування деяких даних. Допустимо, одним з атрибутів клієнта є рівень доходу, який повинен бути представленим у системі одним із значень: дуже низьким, низьким, середнім, високим, дуже високим.

Необхідно визначити градації рівня доходу, у цьому процесі буде потрібно співробітництво аналітика з експертом у предметній області. Можливо, для таких перетворень даних буде потрібно написання спеціальних процедур.

Визначення необхідної кількості даних. При визначенні необхідної кількості даних слід ураховувати, чи є дані впорядкованими чи ні.

Якщо дані впорядковані й ми маємо справу з тимчасовими рядами, бажано знати, чи включає такий набір даних сезонну/циклічну компоненту. У випадку присутності в **наборі** даних сезонної/циклічної компоненти, необхідно мати дані як мінімум за один сезон/цикл.

Якщо дані не впорядковані, тобто події з набору даних не зв'язані за часом, у ході збору даних слід дотримуватися таких правил.

Кількість записів у наборі. Недостатня кількість записів у наборі даних може стати причиною побудови некоректної моделі. З погляду статистики, точність моделі збільшується зі збільшенням кількості досліджуваних даних. Можливо, деякі дані є застарілими або описують якусь нетипову ситуацію, їх потрібно виключити з бази даних. Алгоритми, що використовуються для побудови моделей на надвеликих базах даних, повинні бути масштабованими.

Співвідношення кількості записів у наборі й кількості вхідних змінних. При використанні багатьох алгоритмів необхідно певне (бажане) співвідношення вхідних змінних і кількості спостережень. Кількість записів (прикладів) у наборі даних повинна бути значно більшою кількості факторів (змінних).

Набір даних повинен бути репрезентативним і представляти якнайбільше можливих ситуацій. Пропорції подання різних прикладів у наборі даних повинні відповідати реальній ситуації.

3. *Попередня обробка даних.* Аналізувати можна і якісні, і неякісні дані. Результат буде досягнутий і в тому, і в іншому випадку. Для забезпечення якісного аналізу необхідне проведення попередньої обробки даних, яка є необхідним етапом процесу Data Mining.

Оцінювання якості даних. Дані, отримані в результаті збору, повинні відповідати певним критеріям якості. Таким чином, можна виділити важливий підетап процесу Data Mining – оцінювання якості даних.

Якість даних (Data quality) – це критерій, що визначає повноту, точність, своєчасність і можливість інтерпретації даних.

Дані можуть бути високої якості й низької якості, останні – це так звані брудні або «погані» дані.

Дані високої якості – це повні, точні, своєчасні дані, які піддаються інтерпретації.

Такі дані забезпечують одержання якісного результату: знань, які зможуть підтримувати процес прийняття рішень.

Про важливість обговорюваної проблеми говорить той факт, що «серйозне відношення до якості даних» посідає перше місце серед десяти основних тенденцій, що прогнозуються на початку 2017 року в області Business Intelligence і Сховищ даних компанією Knightsbridge Solutions. Цей прогноз був зроблений в січні 2017 року, а в червні 2017 року Даффі Брансон (Duffie Brunson), один із керівників компанії Knightsbridge Solutions, проаналізував якість даних раніше прогнозів.

Прогноз. Багато компаній стали звертати більше уваги на якість даних, оскільки низька якість коштує грошей у тому розумінні, що веде до зниження продуктивності, прийняттю неправильних бізнес-рішень і неможливості одержати бажаний результат, а також ускладнює виконання вимог законодавства. Тому компанії дійсно мають намір вживати конкретні дії для вирішення проблем якості даних.

Реальність. Дана тенденція зберігається, особливо в індустрії фінансових послуг. У першу чергу це стосується фірм, що намагаються

виконувати угоду Basel II. Неякісні дані не можуть використовуватися в системах оцінки ризиків, які застосовуються для установки цін на кредити й обчислення потреб організації в капіталі. Цікаво відзначити, що суттєво змінилися погляди на способи вирішення проблеми якості даних. Спочатку менеджери звертали основну увагу на інструменти оцінки якості, вважаючи, що «власник» даних повинен вирішувати проблему на рівні джерела, наприклад, очищаючи дані й перенавчаючи співробітників. Але зараз їх погляди суттєво змінилися. Поняття якості даних набагато ширше, чим просто їх акуратне введення в систему на першому етапі. Сьогодні вже є розуміння, що якість даних повинна забезпечуватися процесами витягу, перетворення й завантаження (Extraction, Transformation, Loading – ETL), а також одержання даних із джерел, які готують дані для аналізу.

Розглянемо поняття якості даних більш детально. Дані низької якості, або брудні дані – це відсутні, неточні дані з погляду практичного застосування (наприклад, представлені в невірному форматі, не відповідному до стандарту). Брудні дані з'явилися не сьогодні, вони виникли одночасно із системами введення даних.

Брудні дані можуть з'явитися з різних причин, таким як помилка при введенні даних, використання інших форматів подання або одиниць виміру, невідповідність стандартам, відсутність своєчасного відновлення, невдале відновлення всіх копій даних, невдале видалення записів-дублікатів і т.д.

Необхідно оцінити вартість наявності брудних даних; інакше кажучи, наявність брудних даних може дійсно призвести до фінансових втрат і юридичної відповідальності, якщо їх наявність не запобігається або вони не виявляються й не очищаються.

Описані різні типи брудних даних, серед них виділені такі групи:

- брудні дані, які можуть бути автоматично виявлені й очищені;
- дані, поява яких може бути відвернена;
- дані, які непридатні для автоматичного виявлення й очищення;
- дані, появу яких неможливо запобігти.

Тому важливо розуміти, що спеціальні кошти очищення можуть упоратися не з усіма видами брудних даних.

Найпоширеніші види брудних даних такі:

- пропущені значення;
- дублікати даних;
- шуми й викиди.

Пропущені значення (Missing Values).

Деякі значення даних можуть бути пропущені у зв'язку з тим, що:

- дані взагалі не були зібрані (наприклад, при анкетуванні схований вік);
- деякі атрибути можуть бути незастосовні для деяких об'єктів (наприклад, атрибут «річний дохід» не застосуємо до дитини).

Що можна зробити із пропущеними даними:

- виключити об'єкти із пропущеними значеннями з обробки;
- розрахувати нові значення для пропущених даних;

•ігнорувати пропущені значення в процесі аналізу;

•замінити пропущені значення на можливі значення.

2. Дублювання даних

Набір даних може включати продубльовані дані, тобто **дублікати –** записи з однаковими значеннями всіх атрибутів.

Наявність дублікатів у наборі даних може бути способом підвищення значимості деяких записів. Така необхідність іноді виникає для особливого виділення певних записів з набору даних. Однак у більшості випадків, продубльовані дані є результатом помилок при підготовці даних.

Існує два варіанти обробки дублікатів. При першому варіанті віддаляється вся група записів, що містить дублікати. Цей варіант використовується в тому випадку, якщо наявність дублікатів викликає недовіру до інформації, повністю її знецінює. Другий варіант полягає в заміні групи дублікатів на один унікальний запис.

Шуми й викиди.

Викиди – різко одмінні об’єкти або спостереження в наборі даних.

Шуми й викиди є досить загальною проблемою в аналізі даних. Викиди можуть являти собою окремі спостереження або бути об’єднаними в якісь групи. Завдання аналітика не тільки їх виявити, але й оцінити ступінь їх впливу на результати подальшого аналізу. Якщо викиди є інформативною частиною аналізованого набору даних, використовують робастні методи й процедури.

Досить поширена практика проведення двоетапного аналізу – з викидами та з їхньою відсутністю – і порівняння отриманих результатів.

Різні методи Data Mining мають різну чутливість до викидів, цей факт необхідно враховувати при виборі методу аналізу даних. Також у деякі інструменти Data Mining мають бути вбудовані процедури очищення від шумів і викидів. Візуалізація даних дозволяє представити дані, у тому числі й викиди, у графічному виді. Приклад наявності викидів зображений на діаграмі розсіювання на рис. 8.1, де видні кілька спостережень, що різко відрізняються від інших (спостережень, що перебувають на великій відстані від більшості).

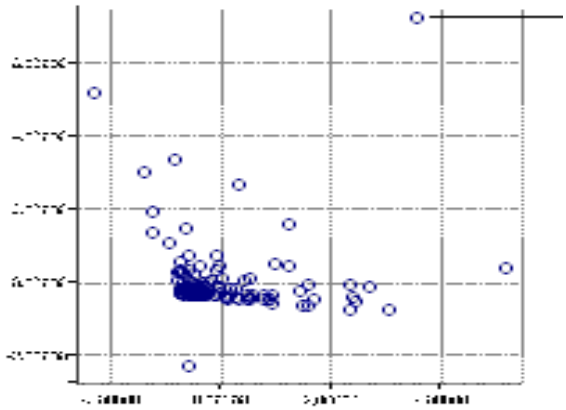

Рисунок 8.1 – Приклад набору даних з викидами

Очевидно, що результати Data Mining на основі брудних даних не можуть вважатися надійними й корисними. Однак наявність таких даних не обов'язково означає необхідність їх очищення або ж запобігання появи. Завжди повинен бути розумний вибір між наявністю брудних даних і вартістю й/або часом, необхідним для їхнього очищення.

3. Очищення даних

Очищення даних (data cleaning, data cleansing або scrubbing) займається виявленням та видаленням помилок і невідповідностей у даних з метою поліпшення якості даних.

Проблеми з якістю зустрічаються в окремих наборах даних – таких як файли й бази даних. Коли інтеграції підлягає безліч джерел даних (наприклад, у Сховищах, інтегрованих системах баз даних або глобальних інформаційних Інтернет-системах), необхідність в очищенні даних суттєво зростає. Це відбувається тому, що джерела часто містять розрізнені дані в різному представленні. Для забезпечення доступу до точних і погоджених даних необхідна консолідація різних представлень даних і виключення інформації, що дублюється. Спеціальні кошти очищення звичайно мають справу з конкретними областями – в основному це імена й адреси – або ж з виключенням дублікатів.

Перетворення забезпечуються або у формі бібліотеки правил, або користувачем в інтерактивному режимі. Перетворення даних можуть бути автоматично отримані за допомогою коштів узгодження схеми.

Метод очищення даних повинен задовольняти таким критеріям:

1. Повинен виявляти й видаляти всі основні помилки й невідповідності і в окремих джерелах даних, і при інтеграції декількох джерел.
2. Метод повинен підтримуватися певними інструментами, щоб скоротити обсяги ручної перевірки й програмування, і бути гнучким у плані роботи з додатковими джерелами.
3. Очищення даних не повинне проводитися у відриві від зв'язаних зі схемою перетворення даних, виконуваних на основі складних метаданих.
4. Функції маппірования для очищення й інших перетворень даних повинні бути визначені декларативним чином та підходити для використання в інших джерелах даних і в обробці запитів.
5. Інфраструктура технологічного процесу повинна особливо інтенсивно підтримуватися для Сховищ даних, забезпечуючи ефективне й надійне виконання всіх етапів перетворення для безлічі джерел і більших наборів даних.

Сьогодні інтерес до очищення даних зростає. Ціла низка дослідницьких груп займається загальними проблемами, пов'язаними з очищенням даних, у тому числі, зі специфічними підходами до Data Mining і перетворенням даних на підставі зіставлення схеми. Останнім часом деякі дослідження торкнулися єдиного, більш складного підходу до очищення рядів даних, який включає аспекти перетворення даних, специфічні оператори та їх реалізації.

4. Етапи очищення даних

Очищення даних включає такі етапи:

1. Аналіз даних.
2. Визначення порядку й правил перетворення даних.
3. Підтвердження.
4. Перетворення.
5. Протитечія очищених даних.

Етап 1. Аналіз даних. Докладний аналіз даних необхідний для виявлення підлягаючих видаленню видів помилок і невідповідностей. Тут можна використовувати і ручну перевірку даних або їх шаблонів, і спеціальні програми для одержання метаданих про властивості даних та визначення проблем якості.

Етап 2. Визначення порядку й правил перетворення даних. Залежно від числа джерел даних, ступені їх неоднорідності й забруднення, дані можуть вимагати досить великого перетворення й очищення. Іноді для відображення джерел загальної моделі даних використовується трансляція схеми; для сховищ даних звичайно використовується реляційне представлення. Перші кроки з очищення можуть уточнити або змінити опис проблем окремих джерел даних, а також підготувати дані для інтеграції. Подальші кроки повинні бути спрямовані на інтеграцію схеми/даних і усунення проблем численних елементів, наприклад, дублікатів. Для сховищ у процесі роботи з визначення ETL повинні бути визначені методи контролю й потік даних, що підлягає перетворенню й очищенню.

Перетворення даних, що зв’язані зі схемою, так само як і етапи очищення, повинні, наскільки можливо, визначатися за допомогою декларативного запиту й мови маппірування, забезпечуючи, таким чином, автоматичну генерацію коду перетворення. До того ж, у процесі перетворення повинна існувати можливість запуску написаного користувачем коду очищення й спеціальних коштів. Етапи перетворення можуть вимагати зворотного зв’язку з користувачем за тими елементами даних, для яких відсутня вбудована логіка очищення.

Етап 3. Підтвердження. На цьому етапі визначається правильність і ефективність процесу перетворення. Це здійснюється шляхом тестування й оцінювання на прикладі або на копії даних джерела, – щоб з’ясувати, чи необхідно якось поліпшити ці визначення. При аналізі, проектуванні й підтвердженні може знадобитися безліч ітерацій, наприклад, у зв’язку з тим, що деякі помилки стають помітні тільки після проведення певних перетворень.

Етап 4. Перетворення. На цьому етапі здійснюється виконання перетворень або в процесі ETL для завантаження й відновлення Сховища даних, або при відповіді на запити по безлічі джерел.

Етап 5. Протитечія очищених даних. Після того, як помилки окремого джерела вилучені, забруднені дані у вихідних джерелах повинні замінитися на очищені, для того, щоб поліпшені дані потрапили також в

успадковані додатки й надалі при витягу не вимагали додаткового очищення. Для Сховищ очищені дані перебувають в області зберігання даних.

Такий процес перетворення вимагає більших обсягів метаданих (схем, характеристик даних рівня схеми, визначень технологічного процесу та ін.). Для погодженості, гнучкості й спрощення використання в інших випадках, ці метадані повинні зберігатися в депозиторії на основі СУБД. Для підтримки якості даних докладна інформація про процес перетворення повинна записуватися як у депозитарій, так і в трансформовані елементи даних (інформація про повноту й актуальність вихідних даних, походження інформації про першоджерело трансформованих об'єктів, зроблені з ними зміни). Наприклад, на мал. 3 похідна таблиця Споживачі містить атрибути Ідентифікатор і Номер, дозволяючи простежити шлях вихідних записів.

Далі докладно описуються можливі методи аналізу даних (виявлення конфліктів), визначення перетворень і розв'язання конфліктів. Конфлікти найменувань звичайно дозволяються шляхом перейменування; структурні конфлікти вимагають часткового перебудування й уніфікації вихідних схем.

Сьогодні ринок програмного забезпечення пропонує великий вибір засобів, метою яких є перетворення й очищення даних.

Розглянемо дві класифікації таких засобів.

Эрхард Рам (Erhard Ram) і Хонг Ганьби До (Hong Hai Do) визначають таку класифікацію засобів очищення й відповідних їм інструментів.

1. Засоби аналізу й модернізації даних.
2. Спеціальні засоби очищення:
   - очищення специфічної області;
   - виключення дублікатів.
3. Інструменти ETL.

1. Засоби аналізу й модернізації даних.

Засоби аналізу й модернізації, що обробляють дані з метою виявлення помилок, невідповідностей і визначення необхідних, що очищають перетворення, згідно із цією класифікацією, можуть бути розділені на засоби профайлинга даних і засоби Data Mining.

Профайлинг даних. MIGRATIONARCHITECT (Evoke Software) є одним із деяких комерційних інструментів цієї категорії. Для кожного атрибута він визначає такі метадані: тип даних, довжину, безліч елементів, дискретні значення та їх процентне відношення, мінімальні й максимальні значення, втрачені значення й унікальність. MIGRATIONARCHITECT також може допомогти в розробці цільової схеми для міграції даних.

Засоби Data Mining. Наприклад, WIZRULE (Wizsoft) і DATAMININGSUITE (Information Discovery) виводять відносини між атрибутами та їх значеннями, обчислюють рівень вірогідності, що відображає число кваліфікуючих рядів.

WIZRULE може відображати три види правил: 1) математичну формулу, 2) правило if-then («якщо-то») і 3) правило правопису. Ці правила відсівають невірно написані імена, наприклад, «значення Edinburgh 52 рази

зустрічається в полі Споживач; у 2-му рядку містять однакові значення». WIZRULE також автоматично вказує на відхилення від набору виявлених правил як на можливі помилки.

Засоби модернізації даних, наприклад, INTEGRITY (Vality), використовують виявлені шаблони й правила для визначення й виконання перетворень, що очищають, тобто модернізують успадковані дані. В INTEGRITY елементи даних зазнають низку обробок – типізація, аналіз шаблонів і частот та ін.

Результатом цих дій є табличне представлення вмісту полів, їхніх шаблонів і частот, залежно від того, які шаблони можна вибрати для стандартизації даних. Для визначення перетворень, що очищають, INTEGRITY пропонує мову з набором операторів для перетворень стовпців (наприклад, переміщення, розщеплення, видалення) і рядків. INTEGRITY ідентифікує й консолідує запис за допомогою методу статистичної відповідності. При обчисленні оцінок для упорядкування відповідностей, за якими користувач відбирає справжні дублікати, використовуються зважені коефіцієнти.

2. Спеціальні засоби очищення.

Спеціальні засоби очищення звичайно мають справу з конкретними областями – в основному це імена й адреси – або ж із виключенням дублікатів. Перетворення або забезпечуються заздалегідь, у формі бібліотеки правил, або в інтерактивному режимі, користувачем. Перетворення даних можуть бути автоматично отримані й за допомогою засобів узгодження схеми.

Низка засобів орієнтована на специфічну область – наприклад, на очищення даних по іменах і адресах або на специфічні фази очищення – наприклад, аналіз даних або виключення дублікатів. Завдяки своїй обмеженості застосування, спеціалізовані засоби звичайно дуже ефективні, однак для роботи із широким спектром проблем перетворення й очищення вони потребують доповнення іншими інструментами.

2.1. Очищення специфічної області.

Імена й адреси записані в різних джерелах і звичайно мають безліч елементів, тому пошук відповідностей їх конкретному споживачеві має велике значення для керування відносинами із клієнтами. Ряд комерційних інструментів, наприклад IDCENTRIC (First Logic), PUREINTEGRATE (Oracle), QUICKADDRESS (QAS Systems),

REUNION (Pitney Bowes) і TRILLIUM (Trillium Software), призначені для очищення саме таких даних. Вони містять відповідні методи: наприклад, метод витягу й перетворення імен і адрес в окремі стандартні елементи, перевірку допустимості назв вулиць, міст і індексів, разом із можливостями зіставлення на основі очищених даних. Вони включають величезну бібліотеку визначених правил щодо проблем, що часто зустрічаються в даних такого роду. Приміром, модуль витяг TRILLIUM і модуль зіставлення містять понад 200000 бізнес-правил. Ці інструменти забезпечують і

можливості настроювання або розширення бібліотеки правил за рахунок правил, визначених користувачем для власних специфічних випадків.

2.2. Виключення дублікатів.

Прикладами засобів для виявлення й видалення дублікатів є DATACLEANSER (EDD), MERGE/PURGELIBRARY (Sagent/Qmsoftware), MATCHIT (Helpitsystems) і MASTERMERGE (Pitney Bowes). Звичайно вони вимагають, щоб джерело даних уже було очищене й підготовлене для узгодження. Ними підтримується кілька підходів до узгодження значень атрибутів; а такі засоби як DATACLEANSER і MERGE/PURGE LIBRARY дозволяють також інтегрувати правила узгодження, визначені користувачем.

3. Інструменти ETL.

Засоби ETL забезпечують можливість складних перетворень і більшої частини технологічного процесу перетворення й очищення даних. Загальною проблемою засобів ETL є обмежені за рахунок власних API і форматів метаданих можливості взаємодії, що ускладнюють спільне використання різних засобів.

Багато комерційних інструментів підтримують процес ETL для Сховищ даних на комплексному рівні, наприклад, COPYMANAGER (Information Builders), DATASTAGE (Informix/Ardent), EXTRACT (ETI), POWERMART (Informatica), DECISIONBASE (CA/Platinum), DATATRANSFORMATIONSERVICE (Microsoft), METASUITE (Minerva/Carleton), SAGENTSOLUTIONPLATFORM (Sagent) і AREHOUSEADMINISTRATOR (SAS). Для однакового керування всіма метаданими по джерелах даних, цільових схемах, маппіруваннях, скриптах і т.д. вони використовують репозиторій на основі СУБД. Схеми й дані вилучаються з оперативних джерел даних як через «рідний» файл і шлюзи СУБД DBMS, так і через стандартні інтерфейси – наприклад, ODBC і EDA. Перетворення даних визначаються через простий графічний інтерфейс. Для визначення індивідуальних кроків маппірування звичайно існує власна мова правил і комплексна бібліотека визначених функцій перетворення. Ці засоби підтримують і повторне використання існуючих перетворених розв’язків, наприклад, зовнішніх процедур C/C++ за допомогою наявного в них інтерфейсу для їхньої інтеграції у внутрішню бібліотеку перетворень. Процес перетворення виконується або системою, що інтерпретує специфічні перетворення в процесі роботи, або відкомпільованим кодом. Усі засоби на базі системи (наприклад, COPYMANAGER, DECISIONBASE, POWERMART, DATASTAGE, WAREHOUSEADMINISTRATOR), мають планувальник і підтримують технологічні процеси зі складними залежностями виконання між етапами перетворення. Технологічний процес може також допомагати роботі зовнішніх засобів (скажімо, у специфічних завданнях очищення це будуть очищення імен/адрес або виключення дублікатів).

Засоби ETL звичайно містять мало вбудованих можливостей очищення, але дозволяють користувачеві визначати функціональність очищення через

власний API. Як правило, аналіз даних для автоматичного виявлення помилок і невідповідностей у даних не підтримується. Проте користувачі можуть реалізовувати таку логіку при роботі з метаданими й шляхом визначення характеристик умісту за допомогою функцій агрегації (sum, count, min, max, median, variance, deviation).

Мови правил звичайно охоплюють конструкції if-then і case, що сприяють обробці виключень у значеннях даних, – невірних написань, абревіатур, втрачених або зашифрованих значень і значень поза припустимим діапазоном. Ці проблеми можуть також вирішуватися за допомогою функціональних можливостей по вибірці даних із таблиць. Підтримка узгодження елементів даних звичайно обмежена використанням можливостей об'єднання й декількох простих строкових функцій відповідності, наприклад точної або групової відповідності або soundex. Проте визначені користувачем функції відповідності полів, так само як і функції кореляції подібності полів, можуть програмуватися й додаватися у внутрішню бібліотеку перетворень.

Інша класифікація засобів очищення даних, запропонована Джулі Борт, підрозділяє інструменти очищення даних на дві умовні категорії:

- універсальні системи, призначені для обслуговування всієї бази даних цілком;
- верифікатори імені/адреси для очищення тільки даних про клієнтів.

Універсальні системи. До цієї категорії належить більша частина продуктів, наявних на ринку. Це: Enterprise Integrator компанії Apertus; Integrity Data Reengineering Tool проведення Validy Technology; Data Quality Administrator від Gladstone Computer Services; Inforefiner фірми Platinium Technology; QDB Analyze (проведення QDB Solutions) Trillium Software System компанії Hart-Hanks Data Technologies.

Ці системи слід вибирати тоді, коли йдеться про створення банків даних усього підприємства й, відповідно, про суцільне очищення даних. Кожна система використовує власну технологію й має власну сферу додатків. Деякі з них працюють у пакетному режимі, наприклад, Trillium, яка переглядає дані в пошуках певних образів і навчається на основі знайденої інформації. Образи, що підлягають розпізнаванню (скажімо, назви фірм або міські адреси), задаються на етапі попереднього програмування. Інші продукти, як то системи компаній Apertus і Validy, являють собою засоби розробки. У першій застосовуються правила, написані мовою Object Query Language. З нею досить легко працювати, але для написання правил потрібна справжня майстерність.

Система компанії Validy при відборі записів використовує алгоритми нечіткої логіки й робить це дуже ефективно, вивуджуючи таке, що людині просто в голову не прийшло б перевіряти. Але цю систему сутужніше освоїти.

Верифікатори імені/адреси. У простих системах, на зразок систем аналізу ринку, цілком можна обійтися очищенням імен і адрес. Приклади

продуктів цієї категорії: Nadis компанії Group 1 Software і пакет компанії Postalsoft. Останній містить три бібліотеки: виправлення й кодування адрес, оформлення правильних імен і злиття/очищення. Перша бібліотека коректує адреси, друга пропонує спосіб їх стандартизації, третя виконує консолідуючі функції.

Ці продукти простіше використовувати, і, оскільки область застосування їх не така широка, роботу з очищення вони виконають значно швидше. Як додаткову функцію це програмне забезпечення надає адресам вид, що відповідає вимогам пошти. Приміром, Nadis автоматично перетворить ім'я й адреса в стандарт Universal Name and Address data standard.

Додатковий продукт компанії Group 1, Code-1 Plus, перевіряє список адрес на відповідність вимогам. Сертифікація гарантує коректність Zip-Коду й використовується при більших обсягах вихідної пошти. Ті, хто застосовував ці засоби, говорять, що автоматизація роботи із забезпечення відповідності адрес різним правилам, установленим поштовим відомством, коштує витрачених зусиль і засобів, навіть якщо доводиться доповнювати названі пакети іншими коштами очищення.

Отже, шляхом використання спеціальних засобів очищення й редагування даних вирішується проблема неякісних або брудних даних. Однак автоматизований процес очищення даних іноді може призводити до помилок у даних, яких раніше в них не було.

Річ Олшефскі (Rich Olshefski) пропонує класифікацію помилок у даних, які виникають у результаті використання засобів очищення. Ці помилки є двома крайностями очищення даних. Якісні, правильно очищені дані перебувають десь на «золотій середині» між цими крайностями по очищенню й редагуванню даних.

Помилка типу 1 виникає, коли інструмент очищення намагається вирішити проблему, якої насправді не існує, тобто починає виправляти невідповідності в даних там, де їх немає.

Помилка типу 2 виникає, коли інструменти очищення повністю упускають існуючу проблему, тобто трапляється при недогляді програмою невірних даних. Такі дані безперешкодно проходять перевірку, будучи при цьому помилковими. Цю помилку ще називають «втраченою помилкою». Програма очищення даних пропускає дані, які насправді повинна була виправити.

Проблема. Саме складне завдання, що постає перед програмою очищення даних, полягає в мінімізації помилок типу 1 і 2. Для усунення помилок Типу 1 програма повинна намагатися не виправляти те, що й так вірно. Це відразу ж закономірним чином підвищує ймовірність виникнення помилки типу 2. Помилок типу 2 можна уникнути шляхом скрупульозної роботи з даними, що, звичайно ж, негайно приводить до зайвого очищення й, відповідно, – до допущення помилки типу 1.

Деякі програми очищення намагаються так чи інакше підтримувати баланс між зайвою старанністю й зайвою довірою, створюючи великі за обсягом звіти про «підозрілі» записи. Ці програми збирають усе підозріле в одну велику купу, яка і є таким звітом. Така методика суттєво збільшує витрати на уточнення даних, оскільки вимагає участі дорогих людських ресурсів.

Іншим шляхом надмірної компенсації помилок типу 1 є внесення занадто малого числа виправлень. А самі примітивні – і тому найнебезпечніші – програми очищення даних намагаються компенсувати й помилки Типу 2, видаючи на виході щось набагато більш кепське, ніж те, що було до «очищення».

Визначення якісної програми очищення даних, за словами Річа Олшефскі, складається із чотирьох елементів. *Програма повинна:*

- не торкатися правильних даних;
- виправляти невірні;
- створювати невеликий за обсягом звіт про підозрілі записи;
- вимагати мінімальних витрат на установку, обслуговування й ручні перевірки.

Питання для самоконтролю

1. Які етапи включає традиційний процес Data Mining?
2. Які кроки та як відбувається постановка задачі Data Mining? Наведіть приклад.
3. Які цілі підготовки даних?
4. Що можна зробити із пропущеними даними?
5. Для чого відбувається дублювання даних?
6. Дайте визначення поняттю «очищення даних» (data cleaning, data cleansing або scrubbing). Для чого робиться очищення даних?