УДК 004.6

В. Яцишин, Ю. Журихін
(Тернопільський національний технічний університет імені Івана Пулюя)

ПРОЦЕСИ ЗАБЕЗПЕЧЕННЯ ЯКОСТІ ДАНИХ ПРИ ПРОЕКТУВАННІ СИСТЕМ МАШИННОГО НАВЧАННЯ

Перш, ніж використовувати дані, потрібно забезпечити їх якість для ефективного використання в системах машинного навчання. Для досягнення цієї мети в рамках однієї системи можуть використовуватися рішення класу Data Quality, а в рамках декількох систем – класу MDM (Master Data Management). Вони дозволяють організувати повний цикл процесів щодо профілізації даних, аналізу їх якості та оптимізації. Реалізація цих процесів призводить до створення еталонних значень. Обробку вихідних даних для приведення їх до еталонних значень можна розбити на ряд процесів: профілювання, стандартизація, очищення, збагачення, дедуплікація.

Процес профілювання – це аналіз існуючих джерел даних з метою визначення їх придатності для використання в планованому бізнес-процесі. Крім того, профілізація дозволяє визначити ті критерії, виконання яких дасть придатні для використання дані. Тобто цей процес допомагає заздалегідь зрозуміти якість і повноту, що міститься в системі інформації для організації нового напрямку.

Наприклад, компанія вирішує організувати розсилання письмових повідомлень своїм клієнтам. Для цього проводиться аналіз їх адрес, в результаті у відсотках оцінюються наявність, реальність адрес і відсутність записів про місце проживання клієнтів. Отримана інформація дозволяє компанії зрозуміти застосовність існуючих даних для організації розсилок.

Процес стандартизації – це приведення даних до єдиного формату. Завданнями стандартизації є нормалізація БД, збільшення атомарності та уніфікація даних. Відзначимо, що кінцева мета нормалізації БД – це зменшення потенційної суперечливості збереженої в базі даних інформації.

Наприклад, у БД компанії інформація про ПІБ клієнтів була введена в різному форматі: у ряді систем ПІБ заносили в одне строкове поле, а в інших для кожного значення була визначена своя колонка в БД. Для стандартизації даних ПІБ, яке вводили в одне поле, слід розбити три поля – на прізвище, ім'я та по батькові.

У свою чергу, уніфікація представлення даних – це процес вибору єдиного формату запису значень. Наприклад, номери телефонів повинні бути приведені до стандартного вигляду, який містить код країни, національний код оператора і номер абонента. Також до стандартизації можна віднести можливість приведення адрес клієнтів до єдиного формату

Процес очищення – це процес виявлення та виправлення помилок і невідповідностей даних. Завдання очищення – аналіз інформації, визначення помилкових даних і усунення неточностей. Типовим випадком останнього є статистичний аналіз даних. Існує безліч методів, які можуть визначати схожість рядків і автоматично виправляти помилки.

Найбільш часто використовуваний в даному випадку алгоритм – обчислення відстані Левенштейна (мінімальна кількість операцій вставки, видалення одного символу та його заміни на інший, необхідних для перетворення одного рядка в інший). Ще одним цікавим прикладом можливості виявлення помилок є метод аналізу контрольних чисел.