System automatycznego zarządzania jakością danych z zastosowaniem metod sztucznej inteligencji i prognozowania

Autor: Szymon Niewiadomski

Data: 18 września 2025

Streszczenie

Bazy danych zarządzania konfiguracją (CMDB) stanowią repozytorium referencyjne dla praktyk ITSM (IT Service Management), w tym między innymi zarządzania zmianą oraz zarządzania pojemnością i wydajnością. Ich jakość danych ma bezpośredni wpływ na ciągłość świadczenia usług i odporność cybernetyczną organizacji. Niniejsza rozprawa przedstawia przejrzystą, modułową i wdrażaną lokalnie (on-premises) architekturę uczenia maszynowego, ukierunkowaną na systematyczne podnoszenie jakości danych CMDB przy zachowaniu wymogów bezpieczeństwa, audytowalności i zgodności charakterystycznych dla operatorów infrastruktury krytycznej. Badania zrealizowano w ramach doktoratu wdrożeniowego we współpracy z przedsiębiorstwem sektora energetycznego, w warunkach, w których przetwarzanie odbywa się lokalnie, modele nie są uprawnione do autonomicznej modyfikacji danych produkcyjnych, a przepustowość weryfikacji eksperckiej jest ograniczona budżetem operacyjnym.

Metodycznie opracowano łańcuch przetwarzania (pipeline) comiesięcznej ewaluacji jakości danych, który *wytypuje* rekordy o podwyższonym ryzyku niezgodności do przeglądu eksperckiego. Atrybuty tekstowe, numeryczne oraz dane słownikowe są reprezentowane w rozbudowanych, automatycznie generowanych przestrzeniach cech (m.in. zliczenia n-gramów, one-hot encoding), po czym stosowana jest selekcja progowa (wariancja/częstość) oraz punktacja dyskryminacyjna oparta na szacowaniu gęstości. Zaproponowano rekurencyjne, jądrowe estymatory gęstości z aktualizacją szerokości jądra i mechanizmem wycofania aktualizacji (rollback), umożliwiające nieparametryczne uczenie w obecności sprzężenia zwrotnego i niepewności etykiet referencyjnych. Dobór cech prowadzony jest z użyciem kryteriów niezależności i zdolności cechy do wykrywania błędu. Wykorzystane są metryki entropowe Rajskiego/Jaccarda. Czterowarstwowy klasyfikator neuronowy łączy informacje dostarczane przez poszczególne cechy tworząc uszeregowaną listę rekordów od najbardziej niepwenego do najbardziej zaufanego. Lista taka może być wykorzystana do ustalenia planu przeglądów i audytu. Uzupełniające moduły wykrywają anomalie numeryczne i strukturalne: metody odległościowe (pełny wariant all-pairs oraz k-NN) oraz kontrole strukturalne wykorzystujące grafową reprezentację CMDB, w tym weryfikację relacji sterowaną identyfikatorami (studium przypadku odwzorowania VIN→producent). Proponowany system implementuje paradygmat nadzoru eksperckiego (human-in-the-loop): każda rekomendacja modelu jest w pełni audytowalna, zweryfikowane werdykty zasilają adaptacyjne aktualizacje, a wolumen i harmonogram inspekcji są dostosowane do dostępnych zasobów.

W ocenie empirycznej modele identyfikatorów osiągają wysoką trafność przy umiarkowanych kosztach obliczeniowych, co umożliwia rzetelne testy spójności strukturalnej. Krzywe precyzja–przepustowość (precision–throughput) potwierdzają skuteczną priorytetyzację rekordów błędnych w zależności od przyjętych progów inspekcji. Eksperymenty pokazują, że łańcuch przetwarzania ujmuje szerokie spektrum niezgodności (duplikaty,

braki identyfikatorów unikatowych, błędy typograficzne, nadużycia pól, niedomknięte procedury operacyjne oraz błędne etykiety cyklu życia) bez polegania na podatnych na erozję zestawach reguł. Założenia architektoniczne—wdrożenie on-premises, niskie wymagania obliczeniowe, wyjaśnialna punktacja i śledzalność offline—wspierają wymagania bezpieczeństwa i zgodności oraz ułatwiają transfer wiedzy pomiędzy pokrewnymi CMDB w obrębie zarządzania jakością danych (data governance).

Drugim wkładem jest ujęcie prognostyczno-optymalizacyjne dla planowania zakupów IT. Rozprawa formułuje zadanie optymalizacji kosztowej z modelami prognostycznymi cen i popytu oraz proponuje algorytm genetyczny (GA) do rozwiązania złożonego, nieliniowego problemu. Zaproponowano unikalną metodę, krzyżowania i mutacji strategi pozyskania zasobów IT. Wyniki eksperymentalne ukazują strategie emergentne (np. konsolidacja zamówień kontra decyzje podejmowane w terminie granicznym) w odmiennych uwarunkowaniach rynkowych i ograniczeniach budżetowo-logistycznych. Przedstawiono również rekomendacje dla monitorowania trendów rynkowych, predykcji czasów dostaw, zmian modeli licencyjnych oraz konsekwencji adopcji usług chmurowych dla polityki zakupowej.

Podsumowując, matematycznie ugruntowane i przejrzyste metody ML pozwalają mierzalnie podnosić jakość danych CMDB oraz wspierać bardziej efektywne planowanie zakupów w reżimie zamówień publicznych.

Słowa kluczowe: ITSM; ITIL; ład danych; zarządzanie jakością danych CMDB; czyszczenie danych; reprezentacja grafowa; informacja wzajemna; odległość Rajskiego; entropia Jaccarda; estymacja gęstości jądrowej; detekcja wartości odstających; interpretowalne sieci neuronowe; prognozowanie popytu i cen; optymalizacja zakupów IT; algorytm genetyczny.