EKSPLORACJA DANYCH W KONTEKŚCIE PROCESU KNOWLEDGE DISCOVERY IN DATABASES (KDD) I METODOLOGII CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

Marcin Mirończuk

Politechnika Białostocka

Eksploracja danych (ED) jest zagadnieniem nowym, mającym niespełna 17 lat. Temat ten jest często poruszany zarówno przez środowiska naukowo-badawcze jak i biznesowe. W przeciągu tego czasu powstało kilka różnych, odrębnych definicji dotyczących kwestii – czym jest i czym się zajmuje Eksploracja Danych (ang. Data Mining). W zależności od przyjmowanego punktu widzenia ED jest traktowana i definiowana jako: nowa metoda naukowa wywodząca się z dziedziny nauki i techniki jaką jest informatyka, pełna i kompletna metodologia CRISP-DM (ang. Cross-Industry Standard Process for Data Mining) służąca do zarządzania projektami Data Mining, lub jako jeden z etapów procesu odkrywania wiedzy z baz danych, określane w skrócie jako KDD (ang. Knowledge Discovery in Databases). Definiowanie ED jako nowej dyscypliny naukowej jest jeszcze zadaniem dość kontrowersyjnym ze względu na niezgodność między badaczami co do dokładnego zakresu i granic swojego obszaru badań. Z tego też względu do Eksploracji Danych aktualnie bezpieczniej jest (do czasu usystematyzowania nauki) odnosić się i osadzać ją w kontekście procesu KDD lub metody CRISP-DM. Knowledge Discovery in Databases jest powszechnie używane w środowisku akademickim naukowo-badawczym natomiast CRISP-DM stworzony został przez koncerny przemysłowe i chętnie adoptowany jest w środowisku biznesowym obok takich rozwiązań jak: Virtuous Cycle of Data Mining, SEMMA (ang. sample, explore, modify, model, assess) czy Six-Sigma.

Referat ma na celu przybliżyć słuchaczom zagadnienia związane z przeprowadzaniem procesu KDD i modelowaniem projektów Data Mining za pomocą CRISP-DM. Przedstawiona zostanie usystematyzowana wiedza, podejścia i pojęcia związane z Data Mining. W pierwszej części referatu zaprezentowane zostanie podejście do Eksploracji Danych jako jednego z cyklu KDD będącego specjalizacją procesu Knowledge Discovery. Następnie zostanie omówiona metoda CRISP-DM. Przytoczony zostanie też kontekst użycia metod w zależności od skali i integracji projektu (zagadnień) którego dotyczą. Na zakończenie nastąpi podsumowanie, w którym wykazane zostaną wspólne cechy między obydwoma podejściami do eksploracji i wydobywania wiedzy z baz danych.