

Przetwarzanie tekstu metodami eksploracji danych

Prof. dr hab. Henryk Rybiński,
Instytut Informatyki Politechniki Warszawskiej
Zakład Systemów Informacyjnych

Streszczenie

Celem wykładu jest przegląd zagadnień związanych z przetwarzaniem danych tekstowych metodami eksploracji danych. W pierwszej kolejności zaprezentowana zostanie dziedzina eksploracji danych tekstowych. Przedstawione zostaną podstawowe pojęcia tej dziedziny, następnie omówione będą podstawowe kierunki badań. Będą też zaprezentowane przykłady zastosowań.

W badaniach odkrywania wiedzy z danych tekstowych wyróżnia się podejścia *knowledge-rich* i *knowledge-poor*. Podejście *knowledge-rich* w wielu przypadkach pozwala uzyskiwać dobre wyniki, jednak podstawową jego wadą jest to, że wymaga ono zewnętrznych bazy wiedzy. Natomiast podejście odkrywanie wiedzy z danych tekstowych metodami eksploracji danych charakteryzuje się tym, że zwykle jest to podejście typu *knowledge-poor*. Zaprezentowane zostaną badania własne bazujące w dużej mierze na podejściu *knowledge-poor*. W szczególności pokazane będą podejścia do odkrywania znaczeń słów w oparciu o metodę zbiorów zamkniętych (SnS). Przedyskutowana zostanie możliwość wykorzystania tej metody w procesie grupowania semantycznego krótkich tekstów. Ponadto będzie pokazany algorytm wspomaganie procesu budowania słowników dwujęzycznych, w którym także możliwe jest wykorzystanie algorytmu SnS do wykrywania tłumaczeń terminów wielojęzycznych.

Na zakończenie pokazane zostaną metody przetwarzania tekstu na potrzeby zrealizowanego w Instytucie Informatyki systemu bazy wiedzy. Przedstawione będą algorytmy wzbogacania semantycznego tekstów naukowych oraz metody wyszukiwania (i rankingu) ekspertów.