

# Nowe podejścia do oceny wartości informacji

Mieczysław Kłopotek  
Instytut Podstaw Informatyki PAN, Warszawa  
[mieczyslaw.klopotek@ipipan.waw.pl](mailto:mieczyslaw.klopotek@ipipan.waw.pl)

## 1. Opis projektu

Celem projektu jest prowadzenie badań w obszarach:

- learning to rank – nauczenie się z interakcji z człowiekiem sposobu rangowania dokumentów w odpowiedzi na zapytanie do wyszukiwarki internetowej;
- sentiment analysis – stwierdzenie zabarwienia emocjonalnego dokumentu (co implikuje subiektywność itp.);
- ekstrakcja hierarchii pojęć z dokumentów – automatyczne tworzenie tezaursów pojęć podobnych, nadrzędnych i podrzędnych w celu wspomagania wyszukiwania faktów i dokumentów;
- wykrywanie spamskich dokumentów, serwisów itd.

W ramach każdego z tych zagadnień celem będzie:

- ewaluacja dotychczasowych modeli;
- opracowanie polepszonych metod i modeli;
- stworzenie zasobów wspomagających realizację danego zadania;
- stworzenie metod i zasobów ewaluacji metod;
- integracja wybranych metod/modeli z systemem semantycznej wyszukiwarki internetowej;
- poszukiwanie efektów synergicznych między poszczególnymi obszarami oraz ich eksploatacja w celach (1) usprawnienia działania wyszukiwarki – jako element badań stosowanych (2) pogłębienia rozumienia konceptu wartości informacji jako element badań podstawowych.

Rozumienie wartości informacji z punktu widzenia ludzkiej percepcji jest nie tylko z teoretycznego lecz także z inżynierskiego punktu widzenia jednym z najistotniejszych konceptów tej części informatyki, która zajmuje się przetwarzaniem informacji w Internecie, ponieważ dotychczasowa interpretacja zawartości informacyjnej, bazująca na entropii Shannona, wydaje się być całkowicie nieadekwatna.

W szczególności palącym problemem jest takie dodefiniowanie informacji, które odpowiadało jej ludzkiemu rozumieniu i pozwalałoby na automatyczną ocenę informacji w sposób zadawalający człowieka. Do elementów takiej oceny należą między innymi badania z takich obszarów jak np. wyżej wymienione learning to rank, sentiment analysis, concept hierarchy extraction czy spam detection.

## 2. Wymagania (oczekiwania)

- a. Ukończone studia drugiego stopnia z informatyki, matematyki lub fizyki

- b. Biegła znajomość języka programowania Java (doświadczenie programistyczne), mile widziana znajomość narzędzi Hadoop i Hive, doświadczenie w pracy w zespołach;
- c. Mile widziane doświadczenie w tworzeniu oprogramowania do przetwarzania wielkich kolekcji dokumentów tekstowych (text i/lub web mining);
- d. Mile widziana znajomość zagadnień analizy tekstów w języku naturalnym
- e. Dobra znajomość języka angielskiego (dobrze widziany certyfikat CAE lub równorzędny)