Rok akademicki 2014/2015

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Informatyki

PRACA DYPLOMOWA MAGISTERSKA

inż. Paweł Szostek

Metadata extraction from scholarly publications

Opiekun pracy
prof. dr hab. inż. Piotr Gawrysiak

Ocena: .....

.....

Podpis Przewodniczącego
Komisji Egzaminu Dyplomowego

Miejsce na zeskanowane lub wklejane zdjęcie

Kierunek: Informatyka

Specjalność: Inżynieria Systemów Informatycznych

Data urodzenia: 1987.11.27

Data rozpoczęcia studiów: 01.10.2010

Życiorys

.....
Podpis studenta

EGZAMIN DYPLOMOWY

Złożył egzamin dyplomowy w dniu .....20__ r

z wynikiem .....

Ogólny wynik studiów: .....

Dodatkowe wnioski i uwagi Komisji: .....

.....

.....

STRESZCZENIE

Managing and processing collections of scientific literature has become an important aspect of digital libraries. Instead of handling publications by hand, publishers try to process the articles automatically in order to extract relevant metadata used to feed publication search engines. In this work we present a machine learning-based system for automatic metadata extraction from scholarly publications. Furthermore, we present a data set of open-access articles, made of full texts and classification data, used for training of the classifiers. Finally, we report validation results comparable to cutting-edge systems currently available.

Keywords: metadata extraction, digital libraries, machine learning, SVM

EXTRAKCJA METADANYCH Z PUBLIKACJI NAUKOWYCH

Zarządzanie i przetwarzania zbiorów literatury naukowej stało się ważnym aspektem bibliotek cyfrowych. Zamiast ręcznie obrabiać publikacje ręcznie, wydawcy próbują przetwarzać artykuły automatycznie w celu wyodrębnienia odpowiednich metadanych stosowanych w wyszukiwarkach publikacji. W tej pracy przedstawiamy system do automatycznej ekstrakcji metadanych z publikacji naukowych oparty o techniki uczenia maszynowego. Ponadto prezentujemy zbiór artykułów do wykorzystania w procesie uczenia klasyfikatorów do przetwarzania publikacji naukowych, składający się z pełnych tekstów opublikowanych na zasadach otwartego dostępu, jak i z danych klasyfikacyjnych. W poniższej pracy przytoczone są także wyniki walidacji systemu, porównywalne do najefektywniejszych systemów obecnie dostępnych na świecie.

Słowa kluczowe: ekstrakcja metadanych, biblioteki cyfrowe, uczenie maszynowe, SVM