Ekstrakcja metadanych z publikacji naukowych

inż. Paweł Szostek

praca była realizowana pod opieką
prof. dr hab. Piotra Gawrysiaka

Instytut Informatyki
Politechnika Warszawska
14.10.2015

Zakres pracy

W ramach pracy magisterskiej:

- zaimplementowałem system do automatycznej ekstrakcji metadanych oparty o maszyny wektorów wspierających (ang. *SVM*)
- stworzyłem zbiór artykułów wolnego dostępu do ćwiczenia systemów do ekstrakcji metadanych i tekstów z publikacji naukowych
- wystroiłem komponenty systemu:
  - algorytm segmentacji tekstu
  - algorytm odtwarzania kolejności czytania (ang. *reading order*)
  - algorytm klasyfikacji

Opisany system był realizowany w ramach pracy w ICM UW.

Motywacja

- Biblioteki cyfrowe zyskały ogromną popularność w świecie naukowym.
- Liczba artykułów tam przechowywanych uniemożliwia przetwarzanie ręczne (IEEE ~4M, Elsevier ~13.5M).
- Zastosowanie bibliotek cyfrowych ewoluowało z miejsca do przechowywania dokumentów, przez silniki do wyszukiwania prac, aż do platform do analizowania współpracy, cytowań i trendów w nauce.
- Artykuły są często dostarczane w formacie PDF - format ten nie przewiduje udostępniania metadanych (czyli *danych o danych*).

Motywacja

- Biblioteki cyfrowe zyskały ogromną popularność w świecie naukowym.
- Liczba artykułów tam przechowywanych uniemożliwia przetwarzanie ręczne (IEEE ~4M, Elsevier ~13.5M).
- Zastosowanie bibliotek cyfrowych ewoluowało z miejsca do przechowywania dokumentów, przez silniki do wyszukiwania prac, aż do platform do analizowania współpracy, cytowań i trendów w nauce.
- Artykuły są często dostarczane w formacie PDF - format ten nie przewiduje udostępniania metadanych (czyli *danych o danych*).

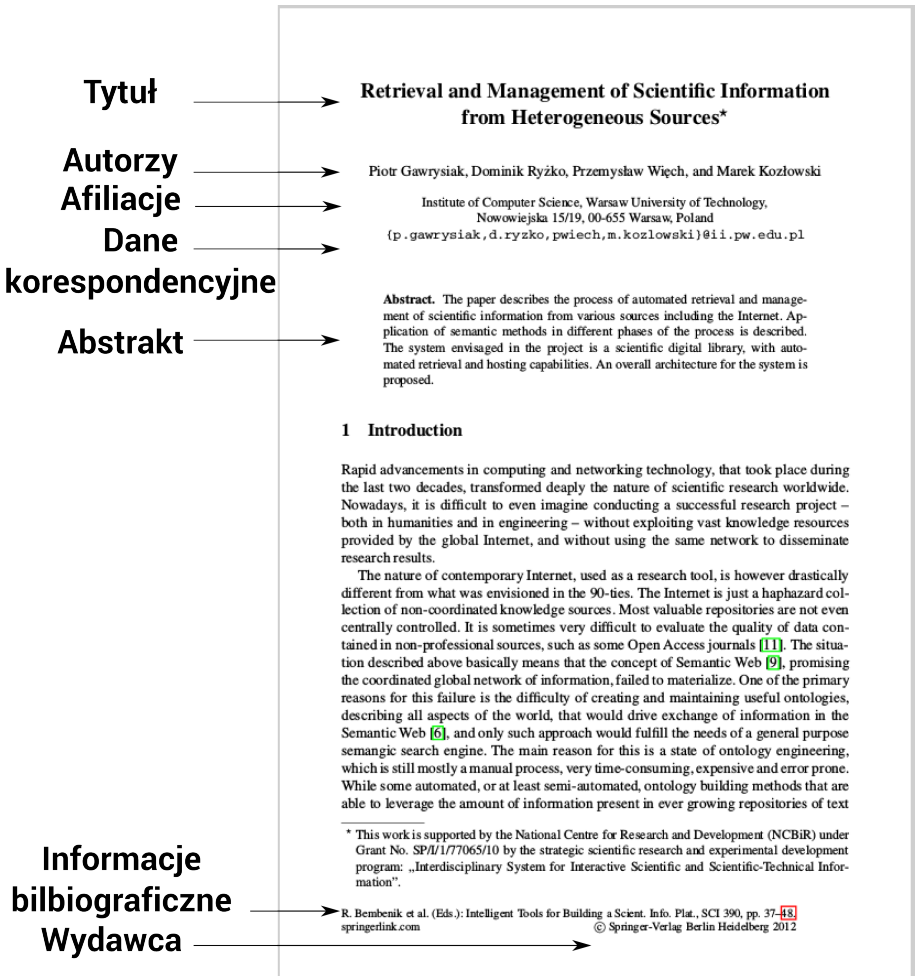
Motywacja

- Biblioteki cyfrowe zyskały ogromną popularność w świecie naukowym.
- Liczba artykułów tam przechowywanych uniemożliwia przetwarzanie ręczne (IEEE ~4M, Elsevier ~13.5M).
- Zastosowanie bibliotek cyfrowych ewoluowało z miejsca do przechowywania dokumentów, przez silniki do wyszukiwania prac, aż do platform do analizowania współpracy, cytowań i trendów w nauce.
- Artykuły są często dostarczane w formacie PDF - format ten nie przewiduje udostępniania metadanych (czyli *danych o danych*).

Motywacja

- Biblioteki cyfrowe zyskały ogromną popularność w świecie naukowym.
- Liczba artykułów tam przechowywanych uniemożliwia przetwarzanie ręczne (IEEE ~4M, Elsevier ~13.5M).
- Zastosowanie bibliotek cyfrowych ewoluowało z miejsca do przechowywania dokumentów, przez silniki do wyszukiwania prac, aż do platform do analizowania współpracy, cytowań i trendów w nauce.
- Artykuły są często dostarczane w formacie PDF - format ten nie przewiduje udostępniania metadanych (czyli *danych o danych*).

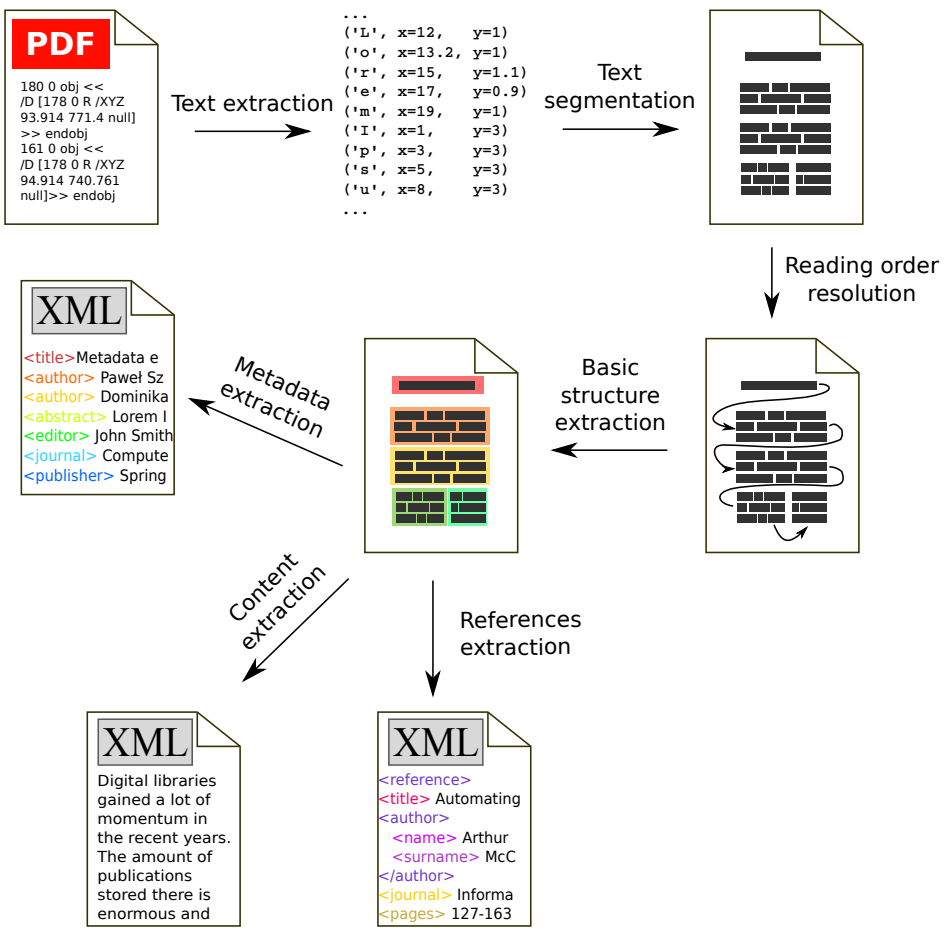
Metadane?



Dodatkowo:

- prawa kopiowania (ang. *copyright*),
- daty dostarczenia tekstu, edycji, publikacji,
- edytor,
- słowa kluczowe,
- typ publikacji,
- referencje bibliograficzne.

Etapy przetwarzania dokumentu



1. ekstrakcja znaków z pliku
2. segmentacja tekstu
3. odtworzenie kolejności czytania
4. wstępna klasyfikacja bloków tesktu
5. przetwarzanie poszczególnych klas
    1. szczegółowa ekstrakcja metadanych
    2. ekstrakcja pełnego tekstu artykułu
    3. ekstrakcja referencji

Maszyny wektorow wspierających (SVM)

- algorytm klasyfikacji (uczenia z nadzorem)
- uczenie polega na znalezieniu optymalnej (hiper-)płaszczyzny separującej zbiór punktów należących do dwóch klas decyzyjnych
- klasyfikacja nieznanych punktów polega na określeniu po której stronie płaszczyzny decyzyjnej znajduje się dany punkt
- jądro przekształcenia pozwala na przeniesienie punktów do przestrzeni o wyższej wymiarowości

- w pracy wykorzystano bibliotekę LibSVM

Maszyny wektorow wspierających (SVM)

- algorytm klasyfikacji (uczenia z nadzorem)
- uczenie polega na znalezieniu optymalnej (hiper-)płaszczyzny separującej zbiór punktów należących do dwóch klas decyzyjnych
- klasyfikacja nieznanych punktów polega na określeniu po której stronie płaszczyzny decyzyjnej znajduje się dany punkt
- jądro przekształcenia pozwala na przeniesienie punktów do przestrzeni o wyższej wymiarowości

- w pracy wykorzystano bibliotekę LibSVM

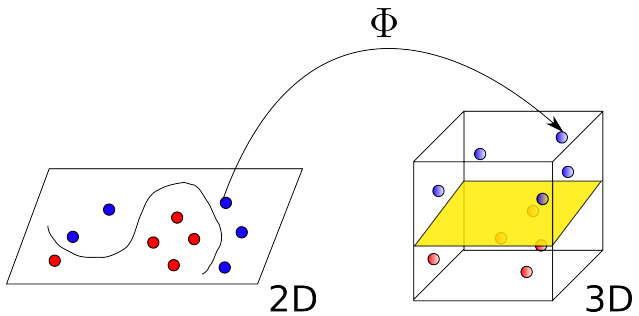
Maszyny wektorow wspierających (SVM)

- algorytm klasyfikacji (uczenia z nadzorem)
- uczenie polega na znalezieniu optymalnej (hiper-)płaszczyzny separującej zbiór punktów należących do dwóch klas decyzyjnych
- klasyfikacja nieznanych punktów polega na określeniu po której stronie płaszczyzny decyzyjnej znajduje się dany punkt
- jądro przekształcenia pozwala na przeniesienie punktów do przestrzeni o wyższej wymiarowości

- w pracy wykorzystano bibliotekę LibSVM

Maszyny wektorow wspierających (SVM)

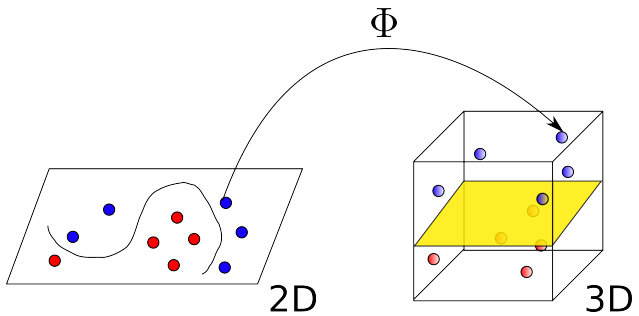
- algorytm klasyfikacji (uczenia z nadzorem)
- uczenie polega na znalezieniu optymalnej (hiper-)płaszczyzny separującej zbiór punktów należących do dwóch klas decyzyjnych
- klasyfikacja nieznanych punktów polega na określeniu po której stronie płaszczyzny decyzyjnej znajduje się dany punkt
- jądro przekształcenia pozwala na przeniesienie punktów do przestrzeni o wyższej wymiarowości



- w pracy wykorzystano bibliotekę LibSVM

Maszyny wektorow wspierających (SVM)

- algorytm klasyfikacji (uczenia z nadzorem)
- uczenie polega na znalezieniu optymalnej (hiper-)płaszczyzny separującej zbiór punktów należących do dwóch klas decyzyjnych
- klasyfikacja nieznanych punktów polega na określeniu po której stronie płaszczyzny decyzyjnej znajduje się dany punkt
- jądro przekształcenia pozwala na przeniesienie punktów do przestrzeni o wyższej wymiarowości



- w pracy wykorzystano bibliotekę LibSVM

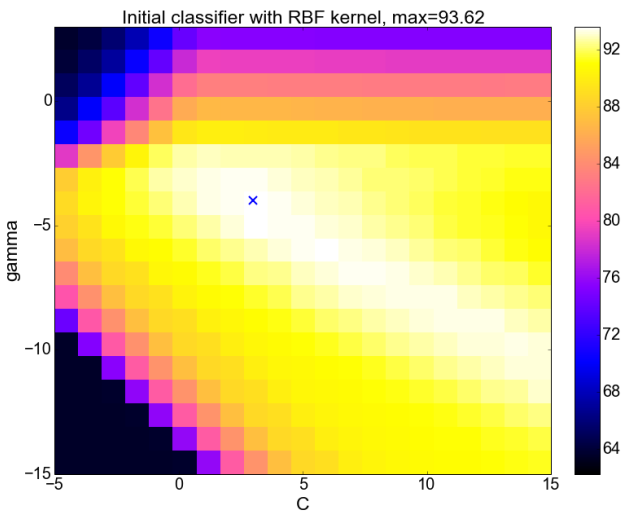
Cechy klasyfikacyjne (ang. *features*) tekstu

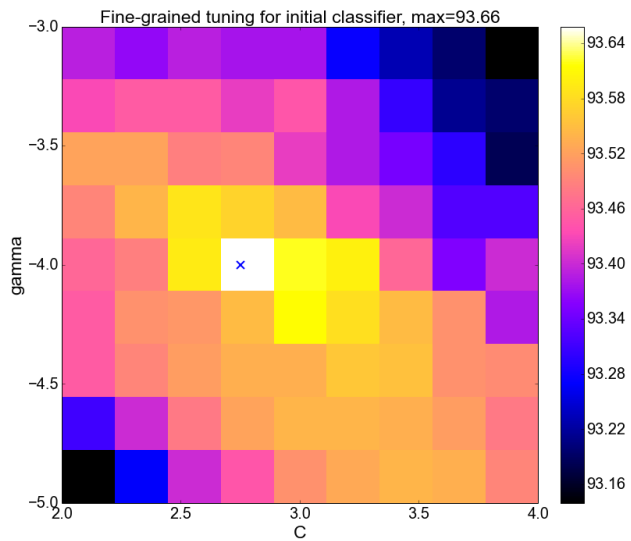
Klasyfikacja tekstu odbywa się per blok. Poniższe cechy są brane pod uwagę:

- Cechy formatowania
  - Cechy użytych fontów, interlinii itp.
- Cechy układu
  - Geometryczne cechy bloków tesktu: szerokość, wysokość, marginesy, odległość od innych bloków itp.
  - DistanceFromNearestNeighbourFeature
- Cechy semantyczne
  - Zawartość predefiniowanych słów kluczowych (*authors*, *email*, *abstract*, *keywords*, etc.)
- Cechy specjalne
  - klasa decyzyjna poprzedniego bloku tekstu
  - *IsAnywhereElseFeature*
  - *BracketCountFeature*
  - *CommaCountFeature*

Strojenie systemu

- System zawiera dwa klasyfikatory SVM, które przeprowadzają klasyfikację wstępną oraz szczegółową klasyfikację metadanych.
- Optymalizacja polega na znalezieniu trójki (K, C, γ), gdzie K to jądro przekształcenia, C to koszt nieprawidłowej klasyfikacji, a γ to parametr przekształcenia.





GROTOAP2 - zbiór artykułów treningowych

- GROTOAP2 jest zbiorem 13,000 artykułów naukowych (pełne teksty w formacie PDF i dane *Ground Truth* w formacie TrueViz) z kolekcji PUBMED. Artykuły pochodzą z 1170 czasopism od 208 wydawców,
- Kolekcja PUBMED zawiera pełne teksty (PDF) oraz metadane (NLM) o różnych stopniu wiarygodności (z reguły zależne od czasopisma),
- Metodologia tworzenia GROTOAP2:
  1. Każdy artykuł z PUBMED został wstępnie przetworzony do formatu TrueViz
  2. Metadane były dopasowywane do bloków tekstu przy użyciu dystansu Watermana-Smitha
  3. Efekty dopasowywania były weryfikowane w procesie iteracyjnym (~40). To pozwoliło na stworzenie reguł heurystycznych poprawiających odkryte nieprawidłowości.

Ewaluacja

Ewaluacja systemu została oparta na 5-krotnej walidacji skrośnej przy użyciu GROTOAP2

	BODY	METADATA	REFERENCES	OTHER	precyzja	kompletność
BODY	**167467**	1394	350	769	96.95	98.52
METADATA	2094	**51707**	52	392	95.99	95.32
REFERENCES	1457	45	**9641**	97	95.67	87.77
OTHER	1724	719	34	**24202**	95.06	90.72

	abstract	affiliation	author	bib_info	copyright	correspondence	dates	editor	keywords	title	title_author	type
precyzja	98.18	95.55	97.71	97.62	95.72	90.04	97.15	98.69	93.67	98.61	96.53	93.73
kompletność	97.96	97.76	94.46	99.05	92.58	89.38	92.16	98.22	80.70	98.74	96.0	87.74

Przyszła praca

Działanie systemu może być poprawione poprzez:

- optymalizacja klasyfikatora poprzez zaaplikowanie zbioru algorytmów (np. WEKA, scikit-learn) i wybór globalnie najefektywniejszego,
- poprawienie pokrycia bloków tekstu w zbiorze GROTOAP2,
- systematyczna optymalizacja algorytmu segmentacji,
- dokładne parsowanie nazwisk autorów (pierwsze i drugie imię, nazwisko),

Demonstracja

Slajdy zapasowe - publikacje

[DSB14] Dominika Tkaczyk, Pawel Szostek, and Lukasz Bolikowski. GROTOAP2 - The methodology for creating a large ground truth dataset of scientific articles. *D-Lib Magazine*, 2014.

[DSF+14] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski. CERMINE - automatic extraction of metadata and references from scientific literature. In *11th IAPR International Workshop on Document Analysis Systems*, pages 11–16, 2014.

[DSF+15] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 2015.