Monitoring polskojęzycznej informacji w Internecie

Wiesław Lubaszewski, Krzysztof Dorosz

Grupa Lingwistyki Komputerowej, AGH
Katedra Lingwistyki Komputerowej, UJ

HLT Days 2012
28 września

Agenda

- Ogólne informacje o zespole i badaniach
- System MPI
- Focused Crawling
- Semantic Driven Crawling
- Podsumowanie

O zespole

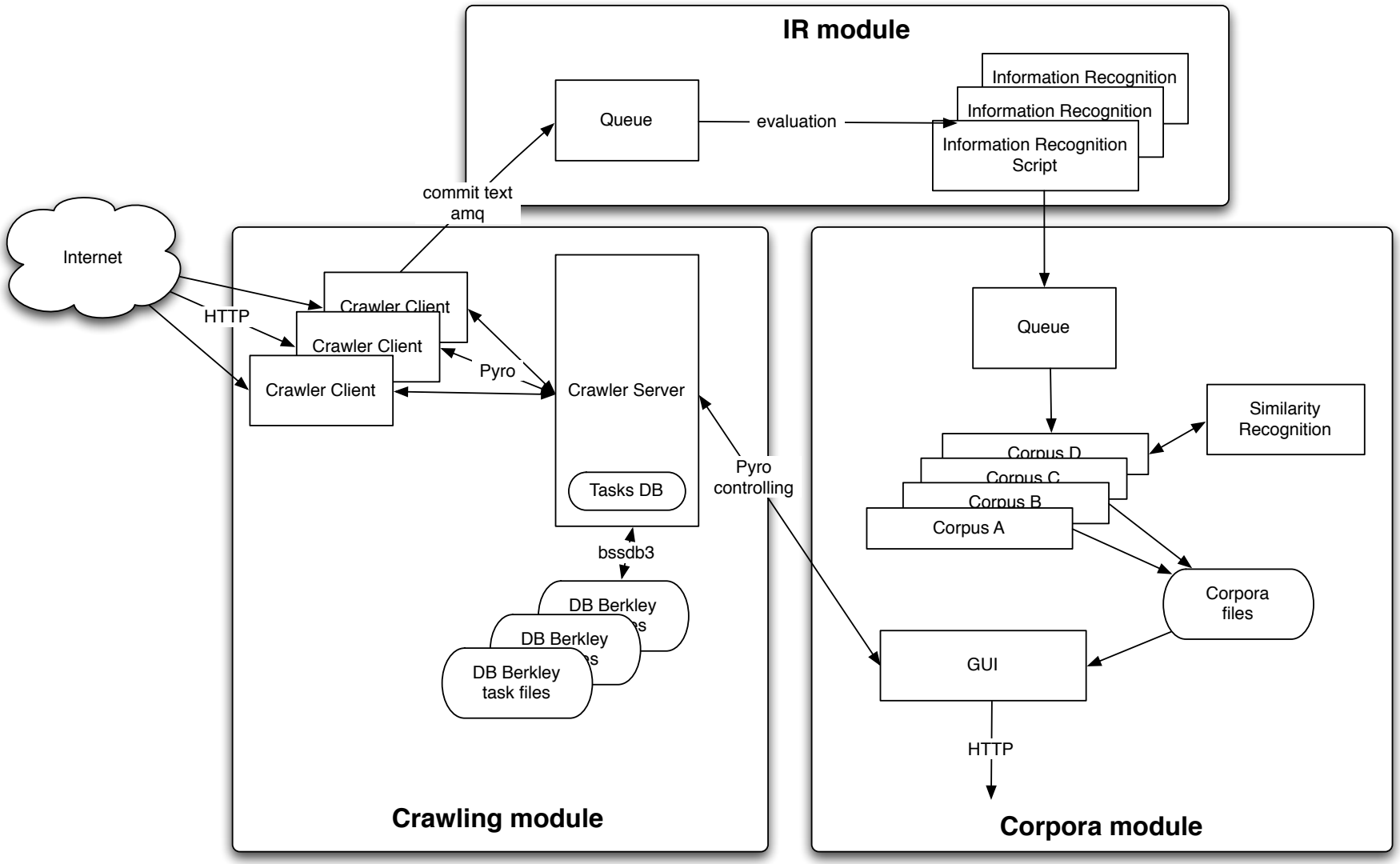
- Międzyuczelniany zespół, na który składa się Grupa Lingwistyki Komputerowej AGH oraz Katedra Lingwistyki Komputerowej UJ.
- Zespół zajmuje się zastosowaniem technik lingwistyczno-komputerowych do przetwarzania tekstu on-line (WWW).
- Słownik Fleksyjny Języka Polskiego - Wydawnictwo Prawnicze LexisNexis, 2001
- Słownik Semantyczny Języka Polskiego
- W ciągu ostatnich 5 lat prac badawczo-rozwojowych został opracowany i rozwijany system MPI - Monitoring Polskojęzycznego Internetu.

Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu / pod red. Wiesława Lubaszewskiego. — Kraków : AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, 2009. — (Wydawnictwa Naukowe / Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie ; KU 0256). — ISBN 978-83-7464-122-7.

System MPI

- Autorska platforma do realizacji zagadnienia przetwarzania tekstów on-line z sieci WWW.
- Składa się z warstwy crawlingowej, metod lingwistycznych, warstwy korpusowej i warstwy prezentacji danych.
- System MPI jest zorientowany na realizację czterech celów:
  A) wyszukiwanie nieznanych stron WWW zawierających określoną treść,
  B) wyszukiwanie informacji w określonej grupie stron WWW,
  C) monitoring zmian informacji na stronach WWW,
  D) wyszukiwanie treści stron WWW podobnych do treści wzorcowej.

Architektura systemu MPI



A) Wyszukiwanie nieznanych stron WWW

- Zadanie polega na odkryciu jak największej liczby nieznanych wcześniej stron WWW zawierających treści pasujące do określonego wzorca semantycznego.
- Wymaga zastosowania specjalizowanego systemu crawlingowego – focused crawling, ze względu na ograniczoną ilość zasobów podczas przeglądania on-line sieci WWW.
- Reprezentacja wzorca informacji jako zdarzenia – model zdarzenia: teoria Conceptual Dependency (Schank, Scripts, Plan Goals ...).
- Problem doboru warunków startowych.
- Ze względu na skalę działania problem jest trudny implementacyjnie.
- Charakter przeszukiwania on-line sieci WWW wymaga skonstruowania nowych miar efektywności.

B) Wyszukiwanie informacji w grupie stron WWW

- Zadanie polega na wyszukaniu precyzyjnej informacji tekstowej (strukturyzowanej) w zakresie znanych stron WWW.
- Możliwe jest zastosowanie prostszych strategii crawlingowych, takich jak klasyczne przeszukiwanie wszerz (BFS).
- Reprezentacja wzorca informacji jako zdarzenia – model zdarzenia: teoria Conceptual Dependency (Schank, Scripts, Plan Goals ...).
- Dużo mniejsza skala przetwarzania; wymaga jednak stosowania dodatkowych technik crawlingu jak omijanie „czarnych dziur”, ukrywanie crawlingu, itp...

C) Monitoring zmian informacji na stronach WWW

- Zadanie polega na detekcji minimalnych zmian treści w wybranym serwisie WWW wykonanych przez człowieka; np. *słońce świeci jasno* $\longrightarrow$ *księżyc świeci jasno*
- Może stanowić narzędzie do wykrywania niejawnej komunikacji pomiędzy wtajemniczoną grupą osób.
- Problem stanowi odróżnienie modyfikacji ręcznej i automatycznej drzewa DOM pomiędzy dwoma pobraniami strony WWW (edycja treści a generowanie treści na stronie).
- Konieczność posiłkowania się mechanizmem regułowym kategoryzującym poszczególne zmiany.

D) Wyszukiwanie treści podobnych

- Zadanie polega na ocenie podobieństwa treści do wzorca reprezentowanego w sposób niestrukturalny, tj. poprzez tekst.
- Wymaga zebrania korpusu tekstów przed przystąpieniem do analizy podobieństwa.
- Podobieństwo ustalane jest poprzez klasyczny LSA (Latent Semantic Analysis), ale prowadzone są także badania nad innymi pokrewnymi metodami asocjacyjnymi.
- Narzędzie to umożliwia kojarzenie tekstów w sposób wykraczający poza rutynę ludzkiego kojarzenia sterowanego przez „principle of the least effort".

Crawling

Web crawling jest to systematyczne gromadzenie treści z sieci World Wide Web z użyciem wyspecjalizowanego narzędzia zdolnego do pracy ciągłej oraz do samoczynnego rozpoznawania topologii połączeń między dokumentami i trawersowania po nich w celu odkrywania treści nieznanych przed rozpoczęciem przetwarzania.

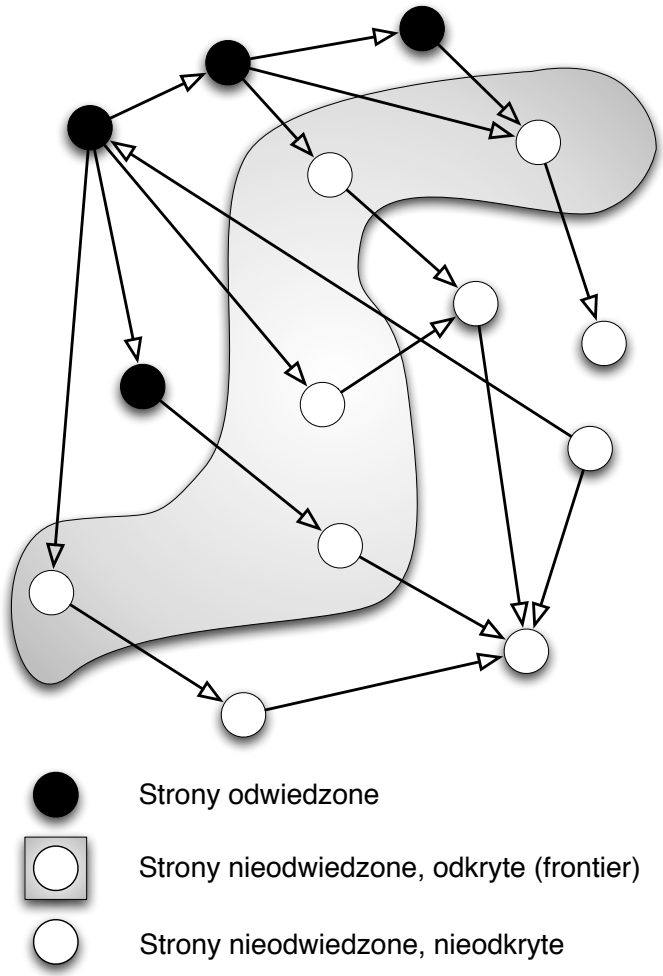
Focused crawling (topic crawling, crawling zorientowany tematycznie):

- pojęcie wprowadzone przez Chakrabarti,
- rodzaj crawlera mającego na celu wyszukiwanie określonych stron, zamiast np. indeksowania,
- wymaga zdefiniowania **wzorca zapytania** T (topic, query) oraz **funkcji podobieństwa** wzorca T do dokumentu P $similarity_T : P \rightarrow v \in [0, 1]$,
- efektywność crawlera to jego zdolności do pobierania stron pasujących do wzorca potencjalnie jak najwcześniej od momentu rozpoczęcia crawlingu.

Strategia crawlingu

Czynniki warunkujące efektywność crawlingu:

- sposób zdefiniowania wzorca zapytania – jakość mechanizmu IR (Information Retrieval),
- specyfika danych i ich rozłożenie w sieci (np. stopień wzajemnego linkowania),
- sposób wyboru warunków początkowych crawlingu (linków startowych),
- **strategia przechodzenia po odsyłaczach (rankingowanie odsyłaczy)**.



Klasyczne strategie crawlingowe

- **Strategia BFS** – oparta na przeszukiwaniu grafu wszerz (Breadth First Search). Odsyłacze odwiedzane są w kolejności ich odkrywania.
- **Strategia DFS** – oparta na wyszukiwaniu w głąb (Depth First Search). Odsyłacze odwiedzane są w natychmiast po ich odkryciu.
- **Strategia Backlink** – rankingowanie oparte na liczbie odsyłaczy linkujących do danej strony. Im więcej odsyłaczy tym szybciej strona zostanie odwiedzona.
- **Strategia Reverse Backlink** – jw., ale im mniej odsyłaczy, tym szybciej strona zostanie odwiedzona.
- **Strategia Random** – strategia referencyjna, polegająca zawsze na wyborze losowego odsyłacza z puli dostępnych.

Semantyczne strategie crawlingowe

- **Strategia Page Value** – rankingowanie odsyłaczy za pomocą wartości oceny semantycznej uzyskanej przez dokument zawierający odsyłacze.
- **Strategia Url Value** – rankingowanie odsyłaczy za pomocą oceny semantycznej treści tekstowej odsyłacza.
- **Strategia Page+Url Value** – kombinacja liniowa dwóch wyżej wymienionych metod.

Wzorzec informacji

Język naturalny dostarcza wielu sposobów opisu zdarzenia. Np. zdarzenie „picia herbaty" może być opisane:

Syntetyczne

Jan pije herbatę.

Analityczne

Promienie słońca rozlały się w pokoju wypełnionym delikatnym aromatem herbaty. Jan wziął powoli pierwszy łyk. Długo delektował się smakiem.

Conceptual Dependency

- Teoria zaproponowana w latach '70 przez R. Schanka.
- Jednym z elementów teorii CD jest kreacja modelu semantycznego opartego na zdarzeniach.
- Realizacją takiego widzenia semantyki jest konstrukcja **skryptu**, stanowiącego **stereotyp zdarzenia** (sekwencji zdarzeń) w, którym występują **role semantyczne**. Role są zmiennymi, które mogą zostać dopasowane w trakcie ewaluacji skryptu.
- Rzeczywistość może zostać opisana w sposób **syntetyczny** lub **analityczny**.

Przykład skryptu CD „picie herbaty”

$ ‘tea drinking’
$Synthetic — *to match event ‘tea drinking’*
((Actor: ?) (Predicate: ‘to drink’, ‘to taste’, ‘to delight in’, ‘to try’, ‘to take a sip of’)
(Object: ‘tea’) (From: ?) (To: ?) (Instrument: ?))
to match a sentence like: X drunk, tasted, sipped some tea, etc.
$Analytic — *to match events: ‘making’, ‘smelling’, ‘operating’, ‘sipping’.*
Making ((Actor: ?) (Predicate: ‘to blend’ ‘to prepare’ ‘to make’) (Object: ‘tea’)
(From: ?) (To: ?) (Instrument: ?)
to match a sentence like: X blended, prepared, made some tea, etc.
Smelling ((Actor: ‘tea’) (Predicate: ‘to smell’ ‘aroma’ ‘fragrance’) (Object: tea)
(Form: ?) (To:?) (Instrument: ?))
to match expressions like: ‘. . . tea smelled, smelt . . . ’ or ‘. . . smell, aroma, fragrance of fresh tea ...’ embedded into a sentence
Operating ((Actor: ?) (Predicate: ‘to reach’, ‘grasp’, ’take off’, ’put on’) (Object:
‘cup’ ‘glass’ ‘tea’) (From: ?) (To: ‘lips’) (Instrument: ‘hand’ ‘fingers’))
to match sentences like: To grasp the cup, To pick up the cup, To take off the cup, etc.
Sipping ((Actor: ?) (Predicate: ‘to take a sip’ ‘to savour a sip’ ‘to delight in a sip’)
(Object: ?) (From: ?) (To: ?) (Instrument: ?))
to match sentences like: X took, savoured, , delighted in a sip, etc.

$End *opracowanie własne*

Proces dopasowywania skryptu CD do tekstu

Proces dopasowania następuje zgodnie z regułą bottom-up i składa się z następujących kroków:

1. **Dopasowanie ról semantycznych** – rola definiowana jest jako lista możliwych elementów (jednostki słownika, wyrażenia wielosegmentowe, cytaty).
2. **Dopasowanie zdarzeń** – realizowane jako dopasowanie zdań skryptu poprzez system wag oraz obligatoryjność ról semantycznych.
3. **Dopasowanie części syntetycznej** – część syntentyczna definiowana jest jako pojedyncze zdanie skryptu.
4. **Dopasowanie części analitycznej** – część analityczna składa się ze zbioru zdań skryptu, dopasowanie realizowane poprzez system wag i obligatoryjność zdań.

Edycja skryptu - Handel organami

Parametry

name: Handel organami

Część syntetyczna | Część analityczna | Część okoliczności

Parametry:

threshold: 55

Zdania

Zdanie 1: Zdanie syntetyczne [weight=20]

Parametry

obligatory: ☑ threshold: 55 name: Zdanie syntetyczne weight: 20

Sloty

sprawca | X

obligatory: ☐ name: sprawca weight: 10 min: 1

zdrowy | X zdrowy | X nałóg | X palący | X palący | X młody | X młody | X młoda | X
młody | X pośrednik | X dawca | X mężczyzna | X kobieta | X Dodaj token

zdarzenie | X

obligatory: ☑ name: zdarzenie weight: 25 min: 1

odstąpić | X kupić | X sprzedać | X sprzedanie | X potrzebować | X potrzebny | X dać | X
oddać | X sprzedaż | X Dodaj token

obiekt | X

obligatory: ☑ name: obiekt weight: 35 min: 1

nerka | X szpik | X wątroba | X śledziona | X narząd | X Dodaj token

cel | X

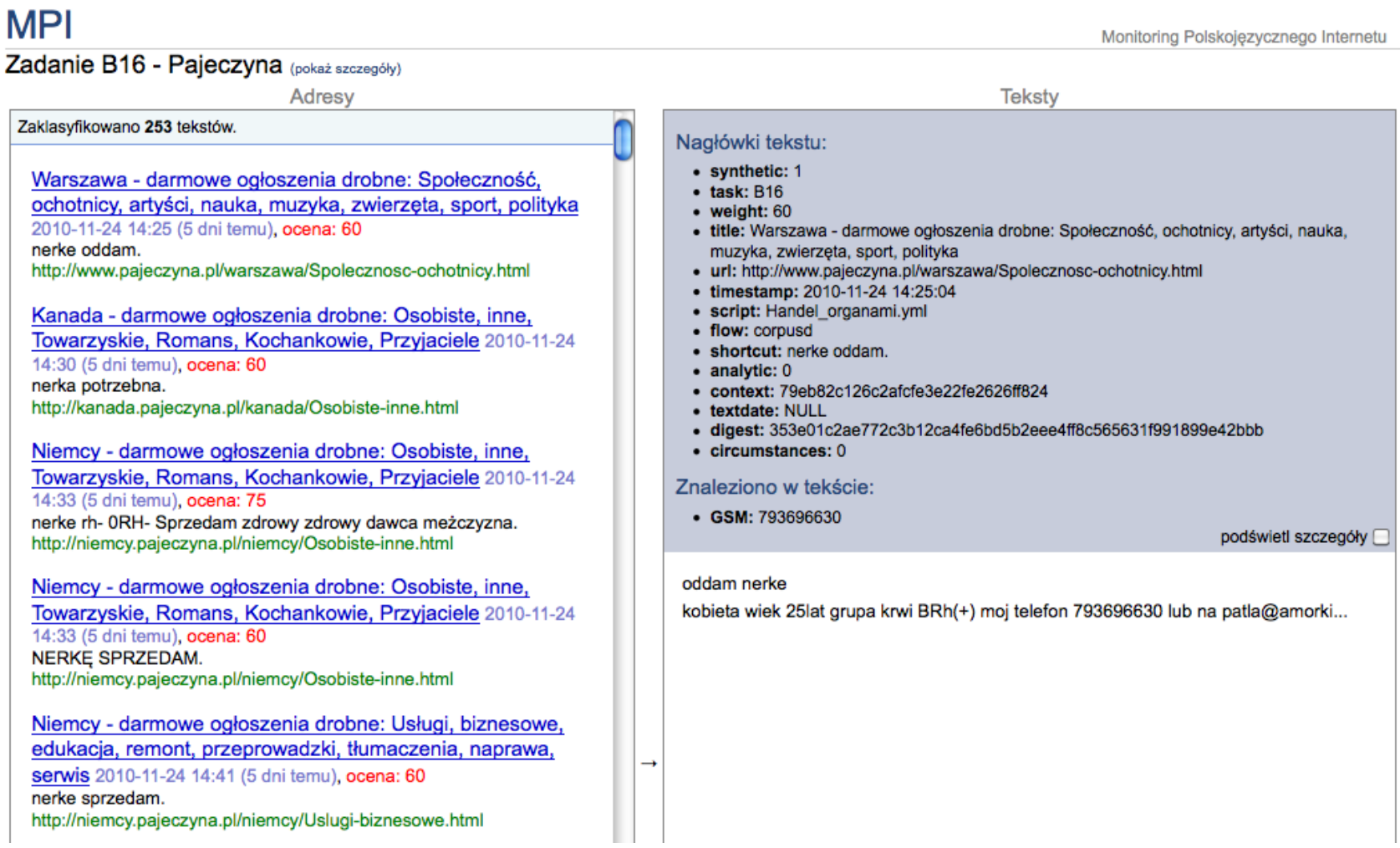
obligatory: ☐ name: cel weight: 25 min: 1

przeszczep | X przeszczepić | X transplantacja | X transplantować | X pośrednictwo | X pomoc | X
pomóc | X Dodaj token

okoliczność | X

obligatory: ☐ name: okoliczność weight: 5 min: 1

tanio | X niedrogo | X pilny | X rh+ | X rh- | X ARH+ | X BRH+ | X ABRH+ | X 0RH+ | X
ARH- | X BRH- | X ABRH- | X 0RH- | X Dodaj token



Dokładność

Uzyskano dla ITA-1: $96,67\%$ oraz dla ITA-2: $81,33\%$

European Seventh Framework Programme
FP7-218086-Collaborative Project

D4.4. System for Enhanced Search:
A Tool for Pattern Based Information Retrieval

The INDECT Consortium

AGH – University of Science and Technology, AGH, Poland
Gdansk University of Technology, GUT, Poland
InnoTec DATA GmbH & Co. KG, INNOTEC, Germany
IP Grenoble (Ensimag), INP, France
MSWiA[1] - General Headquarters of Police (Polish Police), GHP, Poland
Moviquity, MOVIQUITY, Spain
Products and Systems of Information Technology, PSI, Germany
Police Service of Northern Ireland, PSNI, United Kingdom
Poznan University of Technology, PUT, Poland
Universidad Carlos III de Madrid, UC3M, Spain
Technical University of Sofia, TU-SOFIA, Bulgaria
University of Wuppertal, BUW, Germany
University of York, UoY, Great Britain
Technical University of Ostrava, VSB, Czech Republic
Technical University of Kosice, TUKE, Slovakia
X-Art Pro Division G.m.b.H., X-art, Austria
Fachhochschule Technikum Wien, FHTW, Austria

[1]MSWiA (Ministerstwo Spraw Wewnętrznych i Administracji) – Ministry of Interior Affairs and Administration. Polish Police is dependent on the Ministry

Dodatkowe informacje o systemie

D4.4. System for Enhanced Search: A Tool for Pattern Based Information Retrieval

http://www.indect-project.eu/files/deliverables/public/deliverable-4.4

Miary efektywności crawlingu

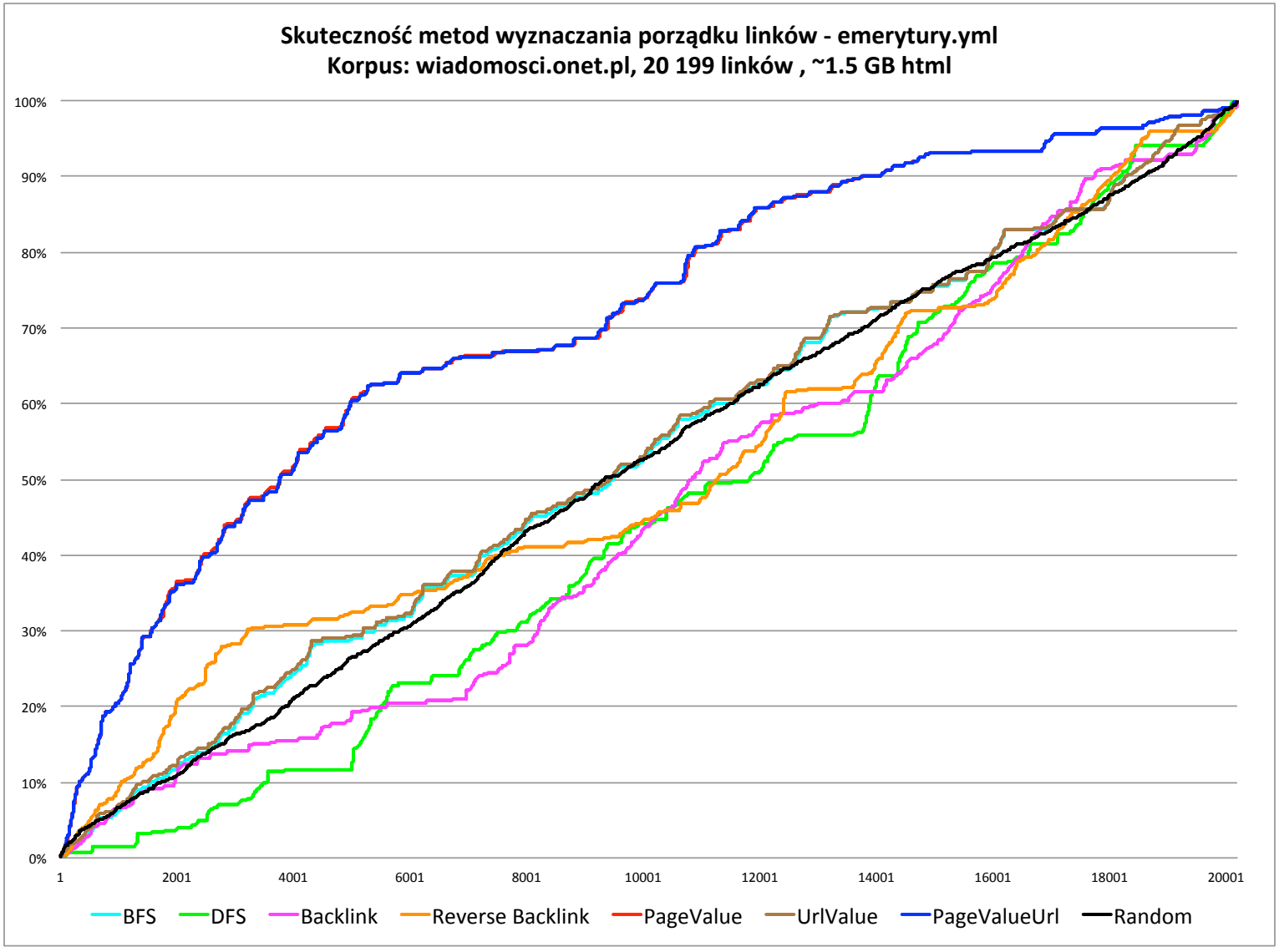
Problem

Użycie typowej miary precision-recall nie jest możliwe, ponieważ recall nie można wyznaczyć w przypadku crawlingu on-line.

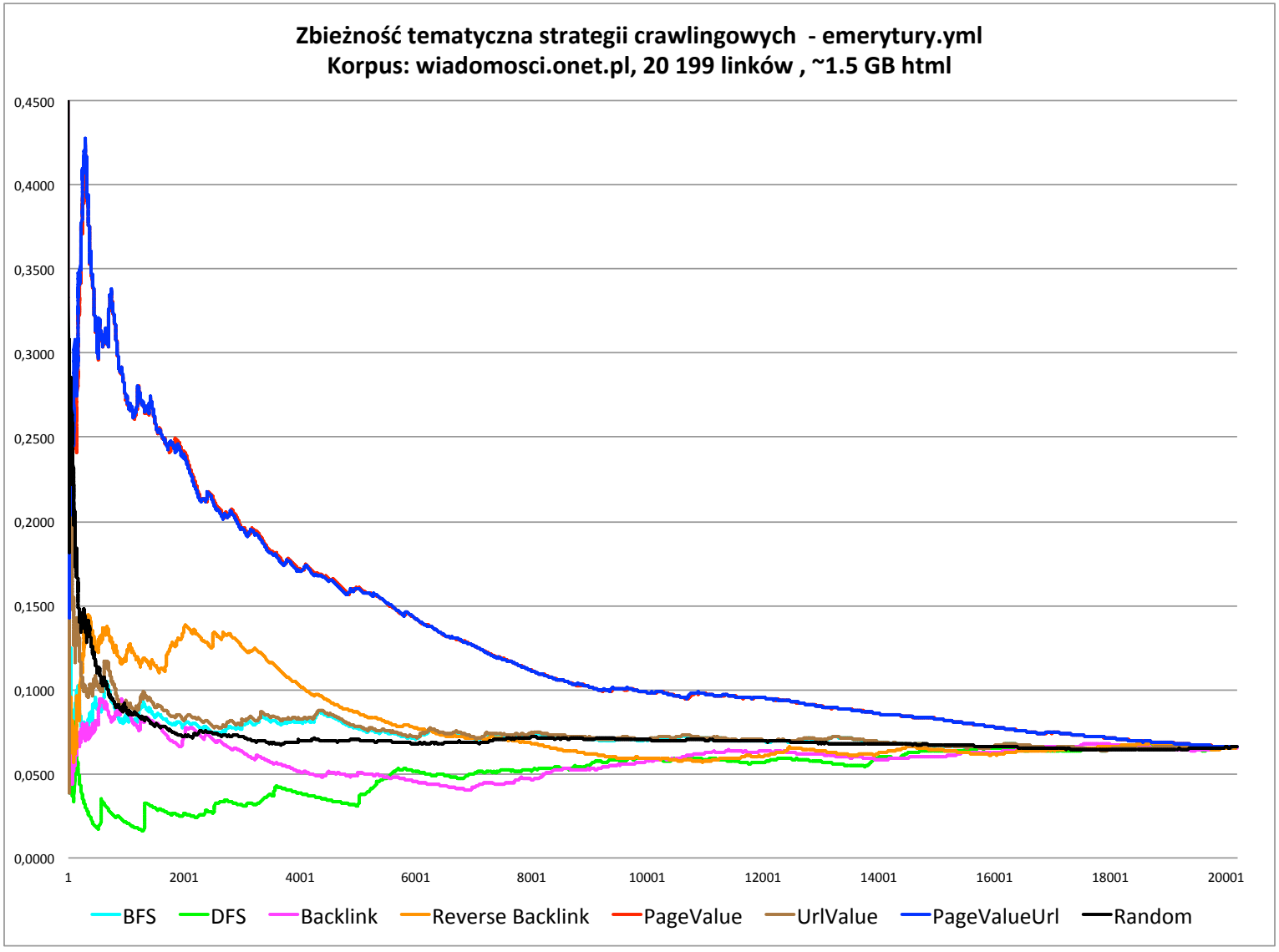
W literaturze można odnaleźć m.in. miary:

- procent odkrytych stron pasujących wzorca w stosunku do czasu,
- moment odkrycia N procent stron pasujących do wzorca,
- zbieżność tematyczna (harvest rate) – $harvest\_rate = \frac{|D_T|}{|D_D|}$, gdzie D_T - zbiór znalezionych dokumentów pasujących do wzorca, D_D - zbiór wszystkich pobranych dokumentów.

Pomiar procentu odkrytych stron



Pomiar zbieżności tematycznej



Podsumowanie

- W 2010 r. Rektor AGH przekazał licencję na prototyp systemu MPI Komendzie Głównej Policji.
- Prace w tym obszarze prowadzone są obecnie w ramach grantu INDECT (7PR).
- Współpracujemy aktywnie z zespołami badawczymi z:
  - University of York, Anglia,
  - Universidad Carlos III de Madrid, Hiszpania.

Dziękuję za uwagę

Wiesław Lubaszewski (lubaszew@agh.edu.pl)
Krzysztof Dorosz (dorosz@agh.edu.pl, krzysztof.dorosz@uj.edu.pl)