Interaktywne wyszukiwanie informacji w repozytoriach danych tekstowych

Marcin Deptuła
Julian Szymański,
Henryk Krawczyk

Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji i Informatyki
Katedra Architektury Systemów Komputerowych

Plan prezentacji

- Motywacja do wykonania badań
- Alternatywny scenariusz wyszukiwania
- Podstawowe składowe algorytmu wyszukującego
- System
- Metody oceny i rezultaty
- Dalsze kierunki badań

Motywacja

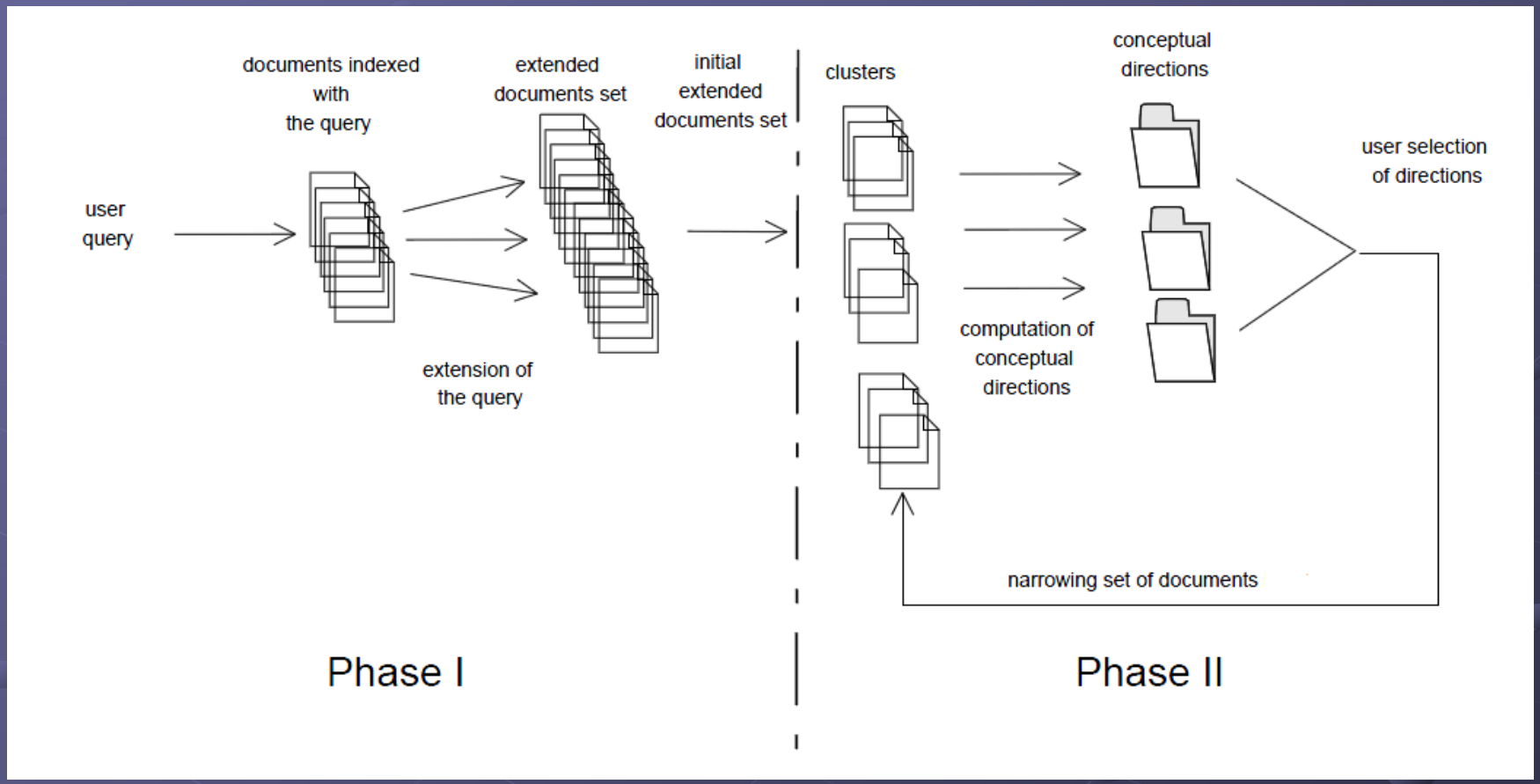
Wyszukiwarki internetowe wykorzystują prosty model komunikacji z użytkownikiem oparty na słowach kluczowych. Pomimo swoich zalet podejście to rodzi kilka problemów.

- Niejednoznaczność języka naturalnego powoduje że te same treści można wyrazić na różny sposób. Również te same słowa mogą opisywać koncepcyjnie różne rzeczy.
- Rezultaty wyszukiwania przeważnie przedstawiane są w postaci posortowanej listy wyników, i nie ma możliwości odsiewania tych rzeczy, które nie pasują do tematyki interesującej użytkownika.
- Wyszukiwanie oparte na słowach kluczowych nie zdaje egzaminu gry użytkownik nie zna słowa kluczowego np.: nazwy leku a potrafi jedynie podać opis np.: wymagań.

Alternatywny scenariusz wyszukiwania

1. Do silnika wyszukującego podane zostają słowa kluczowe
2. Inicjalny zbiór rezultatów zostaje utworzony ze stron zawierających te słowa.
3. W celu poprawy wydajności prezentacji zamiast rankingowanej listy zbiór rezultatów zostaje podzielony na grupy tematyczne.
4. Dla zbioru rezultatów zostają wyznaczone tzw. kierunki konceptualne, które pozwalają różnicować dane.
5. Kierunki konceptualne zostają przedstawione użytkownikowi, tak by mógł on ocenić czy pasują one do tematu wyszukiwania czy nie. Dodatkowo użytkownik może zażądać rozszerzenia jednej ze wskazanych grup tematycznych.
6. W oparciu o interakcję z użytkownikiem zmieniony zostaje zbiór rezultatów końcowych (przez zawężenie lub rozszerzenie)
7. Kontynuowanie wyszukiwania przez przejście do kroku 3

Wyszukiwanie interaktywne



Poprawa miar jakości wyszukiwania

Zwrot

Precyzja

Główne elementy algorytmu

- Warstwa prezentacji – grupowanie rezultatów
  - Zmodyfikowany algorytm klasteryzacji: analiza gęstości i hierarchizacja (DBSCAN + HAC)
- Wyznaczenie kierunków konceptualnych. Największy zysk informacyjny związany z (k_i) kategoriami organizującymi dane

$$IG(k_i) = -\sum_{j=1}^{n} p(w_j) \cdot \log\ p(w_j)$$

$$p(w_j) = \frac{|w_j|}{n}$$

w – waga relacji dokument kategoria

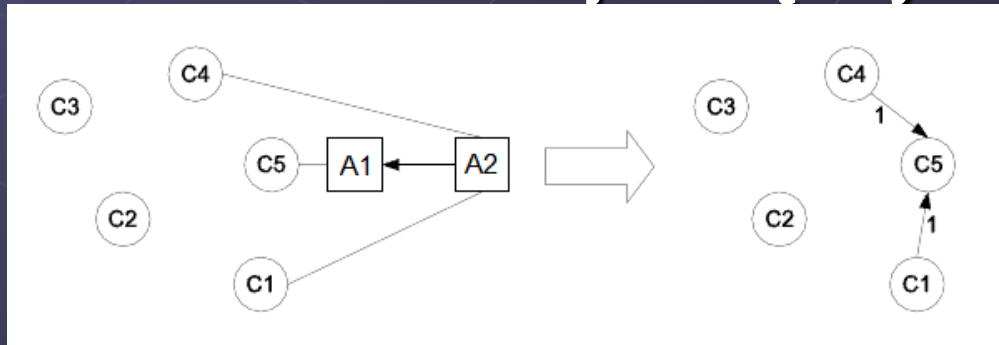
n – liczba unikalnych wartości wagi w kategorii k_i

$|w_j|$ – liczba wystąpień określonej wartości wagi w_j w kategorii k_i

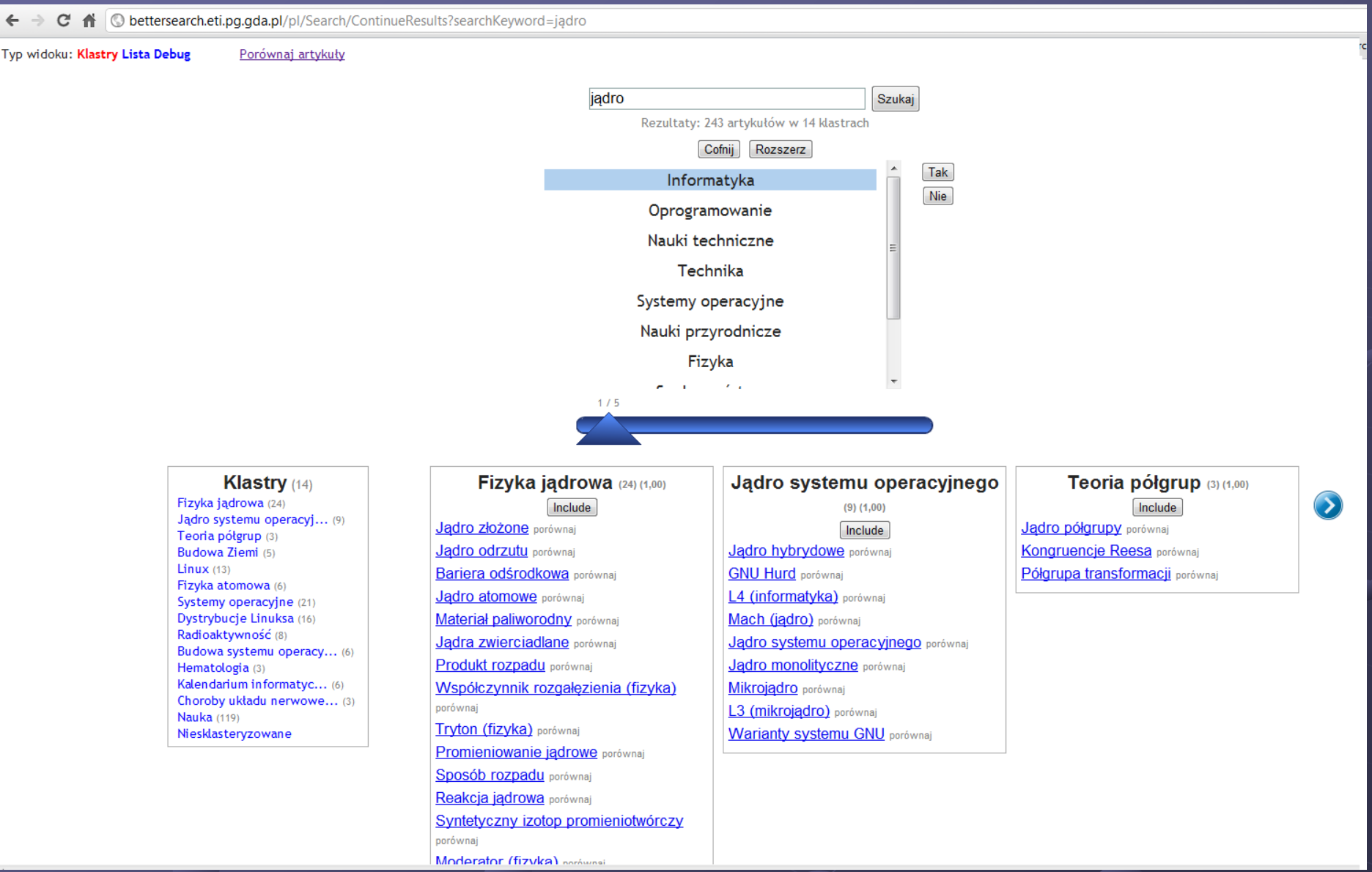
- Rozszerzenie grupy z warstwy prezentacji – dodanie najbardziej zbliżonych dokumentów spoza zbioru rezultatów. Podobieństwo oparte na liczbie powiązań między dokumentami.

Testowe repozytorium

- Zrzuty danych Wikipedii pobrane ze stron Wikimedia foundation.
- Bazy danych dla Polskiej i SimpleEnglish Wikipedii
- Odwrócony indeks do wyszukiwania opartego na słowach kluczowych
- Analiza podobieństwa BOW miedzy dokumentami do wyszukiwania relacji między kategoriami.



http://bettersearch.eti.pg.gda.pl



Metryki oceny

- **Spójność subiektywna**

$$C_c = \frac{|X_c|}{|Y_c|}$$

- X_c – największy zbiór artykułów w grupie które są logicznie powiązane ze sobą (ustalone na podstawie subiektywnej oceny osoby wykonującej test)
- Y_c – zbiór wszystkich artykułów w kastrze
- Spójność dla zapytania wyznaczona jest jako suma spójności we wszystkich grupach podzielona przez ich liczbę.
- Rezultaty dla przykładowych testowych fraz

$$C = \frac{\sum_{c \in C} C_c * |Y_c|}{\sum_{c \in C} |Y_c|}$$

$$C_{min} = \min C_c$$

$$C_{max} = \max C_c$$

Phrase	C	C_{min}	C_{max}	no.ofclusters	no.ofarticles	std.dev.C
Widelec	1.000	1.000	1.000	11	124	0.000
Jądro	0.894	0.600	1.000	26	635	0.1071
Niemcy	0.993	0.857	1.000	11	601	0.0410

Spójność obiektywna

- Pozwala ocenić koncepcyjną spójność klastra bez angażowania czynnika ludzkiego
- Określona jest znormalizowaną minimalną odległością pomiędzy artykułami w grupie a ich kategorią spinającą.

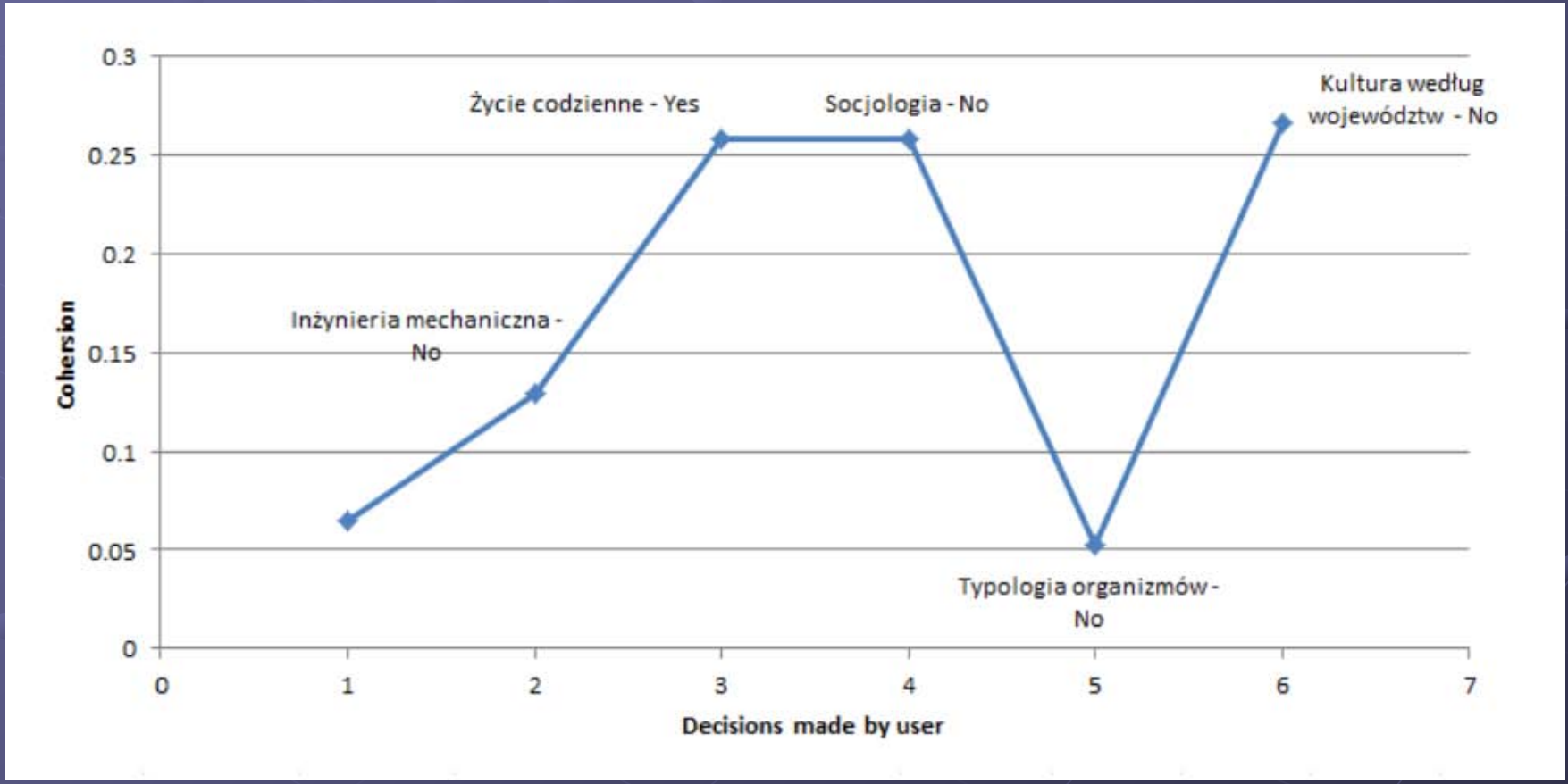
$$O_c = \frac{\sum_{i=1}^{j} k_i}{n}$$

- K_i – liczba artykułów w jednej kategorii, jeśli były one już złączone na kolejnych poziomach hierarchii traktowane są jako jeden dokument.
- j – liczba przejść koniecznych do złączenia dokumentów w jedną kategorię, n – liczba artykułów w grupie

Phrase	AVG(O_c)	no.ofclusters	no.ofarticles
Widelec	1,21	11	124
Jądro	2,02	26	635
Niemcy	1,38	11	601

Phrase	Articles	Minimum grouping category	Distance
Widelec	-Brudny widelec -A teraz coś ... -Latający cyrk Monty Pythona	Monty Python	1.3333
	-Widelec -Łyżkowidelec -Widelczyk -Chochla -Sztućce -Łyżka	Sztućce	1.0000
	-Widelec rowerowy -Amortyzacja rowerowa -Mostek rowerowy -Kierownica rowerowa -Części rowerowe -Rama rowerowa	Części rowerowe	1.0000
	-Piotr Mocarski -Jacek Janowicz -Krzysztof Szubzda	Polscy artyści kabaretowi	1.0000
	-Kabaret Widelec -Gable -Niezbędnik -(63 innych)	Nauka	7.5693
	-Manewr Worka -Belladonna coup	Rozgrywka w brydżu	1.0000
	-Lech -FIS -WFM -MOJ 130 -Perkun -Podkowa -WSK -SHL -WSK M21W2 S2	Polskie motocykle	1.2222
	-BMW R 17 -BMW R-23 -(14 innych)	Motocykle BMW	1.0000
	-Honda CB 600F Hornet -Honda CBR 125R -Honda FMX 650	Motocykle Honda	1.0000
	-Romet 760 -Romet 210 -Komar	Motorowery Romet	1.0000
	-Lista odcinków serial Latający cyrk Montty Pythona -Lista odcinków serialu animowanego Yin Yang Yo! -Lista odcinków serialu Czarodzieje z Waverly Place	Listy odcinków	1.0000

Zbieżność procesu wszukiwania



- Wykres zbieżności spójności w procesie wyszukiwania dla przykładowej frazy widelec (narzędzie).

Dalsze kierunki rozwoju

- Poprawa metod analizy powiązań miedzy kategorialnych
- Rozwój metod wyboru kierunków konceptualnych
- Rozszerzenie wyszukiwania opartego na słowach kluczowych poprzez model HAL analizującego współwystępowanie słów w dużym korpusie tekstowym.
- Poprawa wydajności -> English Wikipedia
- Rozwinięcie metod oceny – Utworzenie zbiorów rezultatów istotnych dla wyznaczania metryk precyzji i zwrotu.
- Klasyfikator wyników wyszukiwarek internetowych

Dziękuję za uwagę