Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych

Piotr Borkowski

Zespół Podstaw Sztucznej Inteligencji
IPI PAN

21 maja 2018

Kategoryzacja dokumentów tekstowych

- Kategoryzacja dokumentów tekstowych polega na przypisaniu jednej lub wielu etykiet (kategorii) do dokumentu tekstowego.
- W tym przepadku proces jest nienadzorowany, tzn. działa bez specjalnie przygotowanego zbioru uczącego.
- Bazuje na informacji semantycznej pochodzącej z już istniejących zasobów w postaci struktury taksonomii np. Wikipedia, MeSH.
- Proces kategoryzacji bazuje tylko na strukturze grafu kategorii oraz nazwach przynależnych do niego konceptów (stron).
- Metoda jest punktem wyjścia do konstrukcji klasyfikatora semantycznego oraz heterogenicznego komitetu klasyfikatorów; metody te sprawdzają się dobrze dla danych z luką semantyczną.

Algorytm kategoryzacji z użyciem taksonomii pojęć

Na wejściu potrzebna jest struktura **abstrakcyjnej taksonomii** (zbiór konceptów i graf kategorii). Kategoryzacja przebiega następująco:

1. dla słów i fraz z dokumentu obliczane są wagi *tfidf* (potrzebny jest korpus dokumentów), wagi te są osobno normalizowane,
2. słowa i frazy odwzorowywane są na koncepty (np. strony Wikipedii, koncepty związane z obiektami MESH),
3. dysponując zbiorem konceptów otrzymuje się zbiór kategorii (np. kategorie Wikipedii, obiekty hierarchii MESH), do których koncepty te są przypisane,
4. dokonuje się agregacji i ustala najczęściej powtarzające się kategorie,
5. stosuje się rzutowanie, jeśli wynik pochodzić ma z określonego zbioru kategorii.

Miary semantycznego podobieństwa

Definiowane miary podobieństwa wykorzystywane są przy ujednoznacznianiu konceptów, agregacji oraz do oceny jakości uzyskiwanych wyników.
Do struktury zakładanej taksonomii zaadaptowano miary dla kategorii znane z literatury poświęconej Wordnetowi:

- Miara *Information Content (IC)* dla kategorii k:

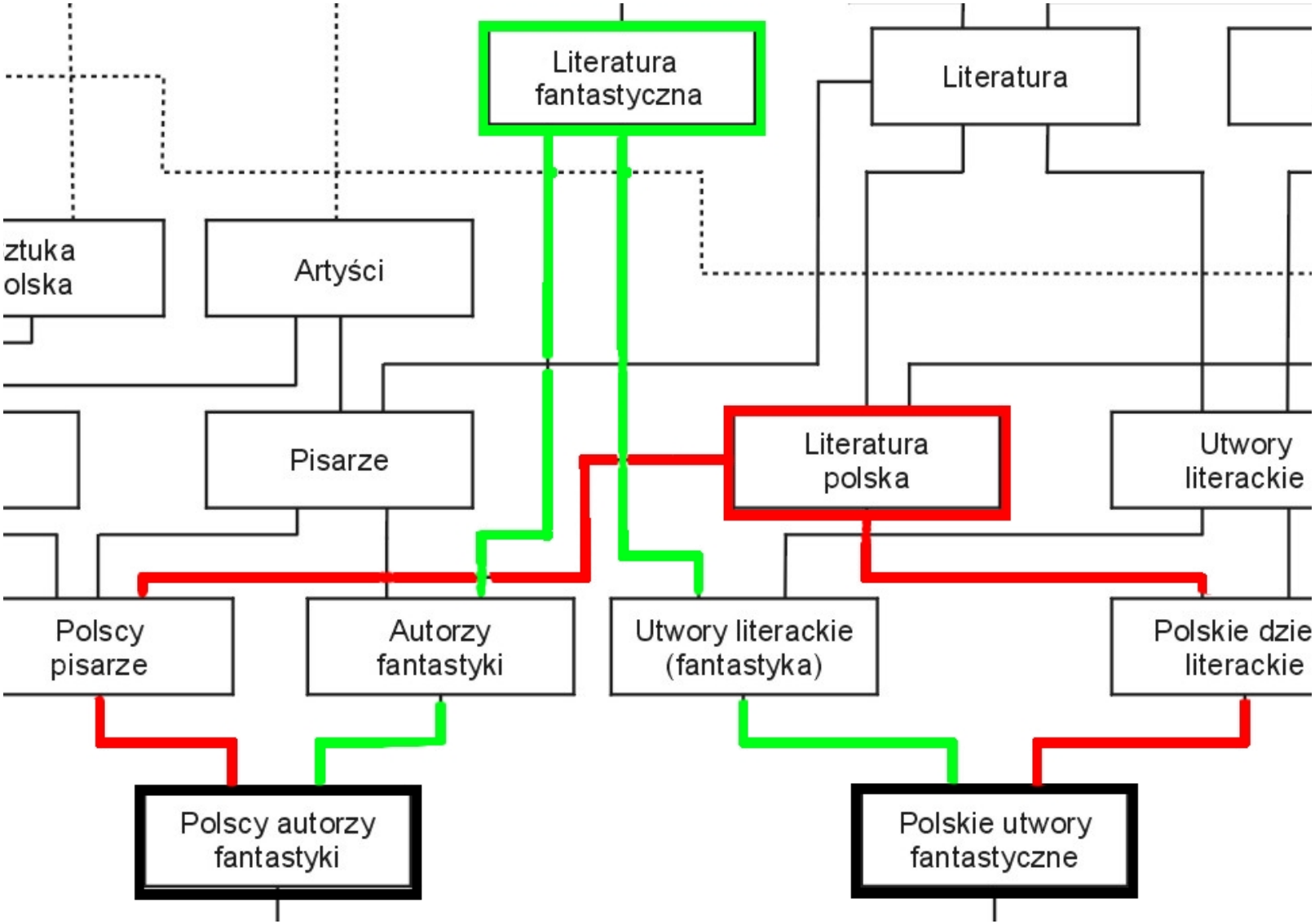
$$IC(k) = 1 - \frac{log\,(1 + s_k)}{log\,(1 + N)},$$

  gdzie s_k jest liczbą konceptów przynależnych do danej kategorii i jej podkategorii, N jest liczbą konceptów w taksonomii.
- Dla zadanych kategorii k_1 oraz k_2 definiujemy:

$$MSCAIC(k_1, k_2) = \max\{IC(k) : k \in CA\},$$

  gdzie CA jest zbiorem kategorii nadrzędnych zarówno w stosunku do kategorii k_1, jak i k_2, oraz:
- $MSCA(k_1, k_2) = \left\{k \in K : IC(k) = MSCAIC(k_1, k_2)\right\}$

Kategorie Wikipedii: **M**ost **S**pecific **C**ommon **A**bstraction



Miary podobieństwa kategorii i stron

- Miary podobieństwa kategorii wzorowane są na miarach stosowanych dla Wordnetu (praca *Pirro-Seco, 2008*). Definiuje się je następująco:

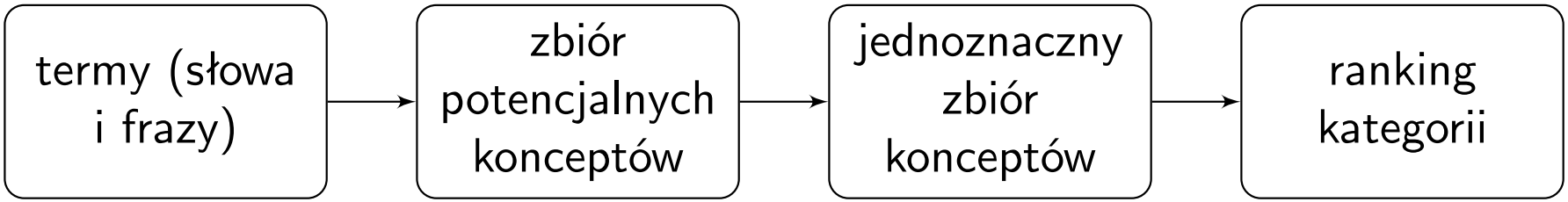
  1. $sim_{\mathrm{Lin}}(k_1, k_2) = \frac{2 \cdot MSCAIC(k_1, k_2)}{IC(k_1) + IC(k_2)}$,
  2. $sim_{\mathrm{Pirro-Seco}}(k_1, k_2) = \frac{1}{3}\Big(2 + 3 \cdot MSCAIC(k_1, k_2) - IC(k_1) - IC(k_2)\Big)$,

- Korzystając z podobieństwa pomiędzy kategoriami zdefiniowano podobieństwo pomiędzy konceptami (stronami), dla p_i i p_j mamy:

$$sim_{\mathrm{KONCEPT}}(p_i, p_j) = max\{sim_{\mathrm{KATEGORIE}}(k_i, k_j) : p_i \in k_i \wedge p_j \in k_j\}$$

Ujednoznacznianie

waga(koncept) $\rightarrow$ waga(kategoria)

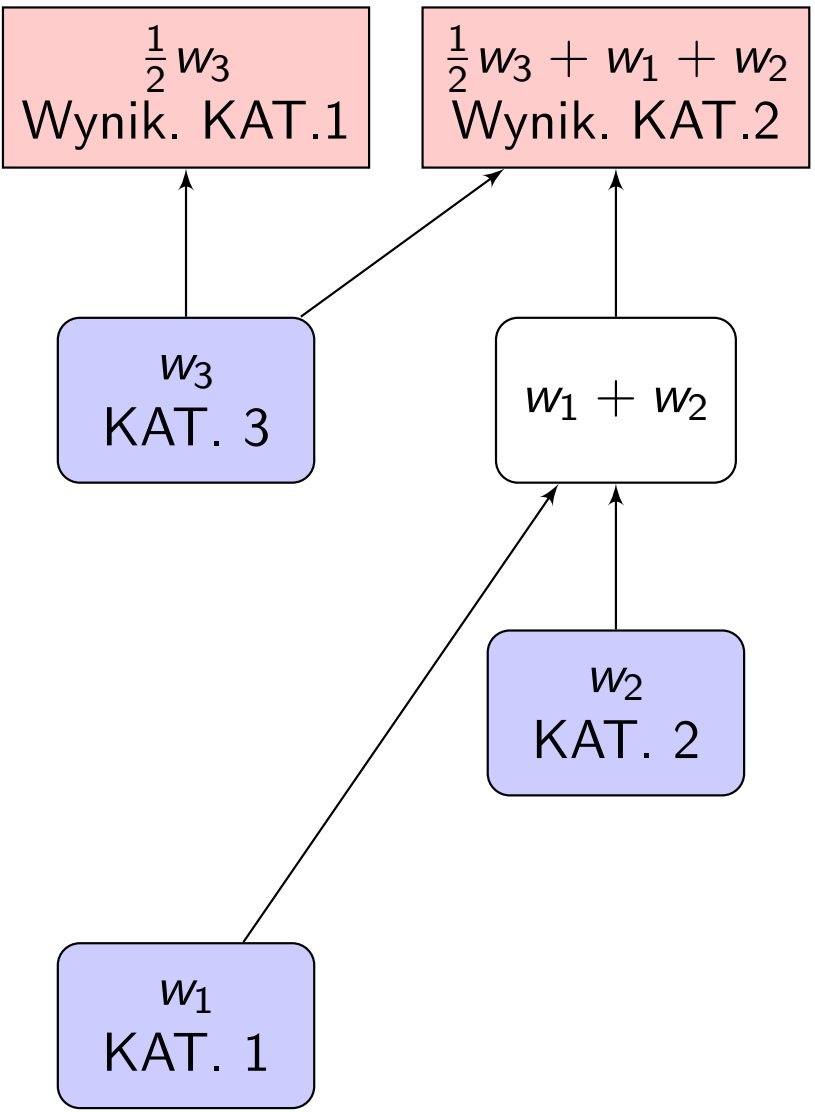


tfidf(t) $\rightarrow$ waga(koncept)

W trakcie mapowania słów z dokumentu na koncepty mogą wystąpić niejednoznaczności, dlatego w jego trakcie wykorzystywana jest metoda ujednoznaczniająca.

Proces ujednoznacznienia słów wieloznacznych bazuje na obliczeniu ich odległości (wg miary podobieństwa semantycznego) do zbioru konceptów jednoznacznych. Znaczenie (koncept), które znajduje się najbliżej, ustalane jest jako ujednoznaczniające dany term.

Rzutowanie kategorii



Szczegółowo zbadano różne warianty algorytmów rzutujących kategorie, spośród których można wyróżnić dwa główne rodzaje:

1. Algorytm uogólniania początkowego rankingu kategorii $k_1, k_2, \ldots, k_R$ poprzez transformację ich do zadanego (przez użytkownika) zbioru kategorii wynikowych $l_1, l_2, \ldots, l_T$.
2. Nienadzorowany algorytm (**agregacji adaptacyjnej**) będący rozwinięciem powyższego algorytmu. Zbiór kategorii wynikowych $l_1, l_2, \ldots, l_T$ jest konstruowany w sposób nienadzorowany, który bazuje na kategoriach wejściowych podlegających agregacji.

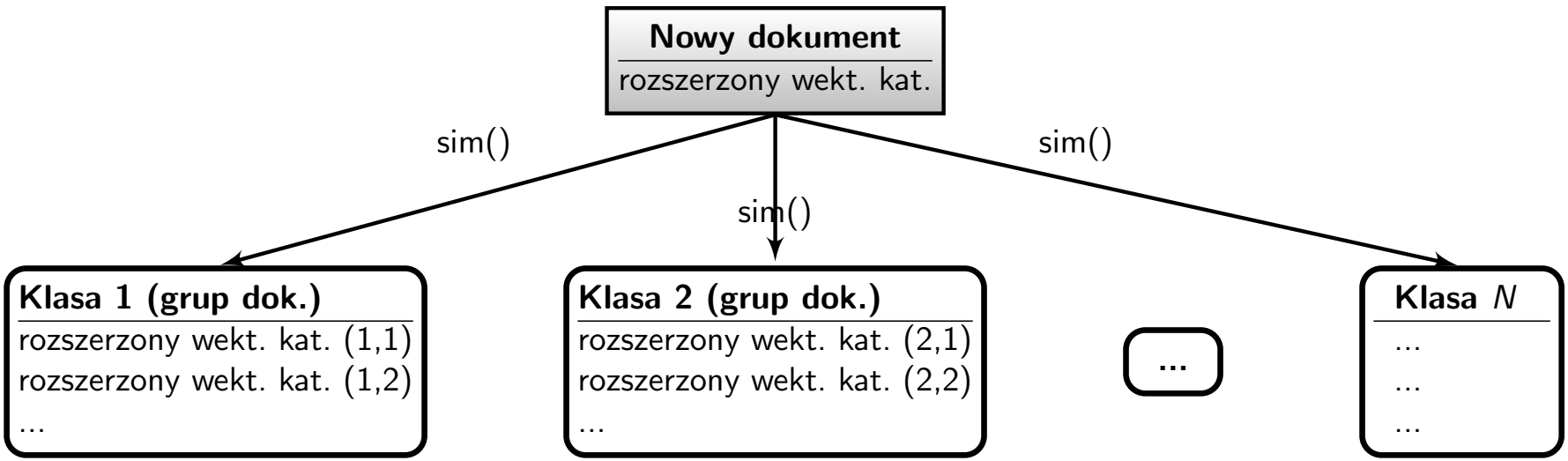
Eksperymenty dla rzutowania kategorii

	precyzja Lin			Shortest path			std. precyzja		
	krótkie	średnie	długie	krótkie	średnie	długie	krótkie	średnie	długie
korpus DMOZ z 15 kategoriami									
bez agregacji	0,143 (0,191)	0,179 (0,217)	0,234 (0,26)	0,222 (0,179)	0,263 (0,188)	0,3 (0,224)	0 (0)	0,005 (0,067)	0,023 (0,152)
agreg. adapt. $T = 1$	0,261 (0,297)	0,336 (0,338)	0,403 (0,342)	0,293 (0,229)	0,334 (0,267)	0,364 (0,252)	0,044 (0,206)	0,089 (0,285)	0,063 (0,243)
agreg. adapt. $T = 5$	0,254 (0,314)	0,334 (0,346)	0,383 (0,355)	0,306 (0,255)	0,358 (0,284)	0,391 (0,285)	0,063 (0,243)	0,103 (0,304)	0,109 (0,313)
agreg. adapt. $T = \infty$	0,221 (0,296)	0,277 (0,328)	0,301 (0,314)	0,314 (0,238)	0,356 (0,26)	0,357 (0,239)	0,04 (0,197)	0,072 (0,259)	0,039 (0,195)
rzutow. na zbiór	0,496 (0,465)	0,598 (0,459)	0,681 (0,426)	0,578 (0,373)	0,665 (0,368)	0,724 (0,349)	0,423 (0,495)	0,535 (0,499)	0,602 (0,492)

Tabela : Wartości średnie dla różnych miar precyzji dla dokumentów z korpusu *DMOZ z 15 kategoriami* przy zastosowaniu różnych metod agregacji

Klasyfikator semantyczny

W ramach prac badawczych powstała metoda klasyfikacji nazwana **klasyfikatorem semantycznym**. Używa ona kategoryjnej reprezentacji dokumentu uzyskiwanej w wyniku działania metody kategoryzującej.



Opracowano też **heterogeniczny komitet klasyfikatorów z dołączoną metodą kategoryzującą**, który pozwala stosować standardowe klasyfikatory w połączeniu z klasyfikatorem semantycznym.

	miara precyzji Lin				miara precyzji „1-0”			
	krótkie	średnie	długie	całość	krótkie	średnie	długie	całość
SEMCAT (1 przebieg)	0,496 (0,465)	0,598 (0,459)	0,681 (0,426)	0,58 (0,46)	0,423 (0,495)	0,535 (0,499)	0,602 (0,492)	0,513 (0,5)
Klasyfikatory – średnie z 25 przebiegów, próba ucząca z Wikipedii L=2 S=200								
Bal. Winn.	0,565 (0,025)	0,68 (0,028)	0,752 (0,034)	0,654 (0,023)	0,5 (0,025)	0,624 (0,03)	0,699 (0,035)	0,599 (0,024)
Bayes	0,459 (0,03)	0,548 (0,024)	0,679 (0,026)	0,533 (0,023)	0,373 (0,045)	0,473 (0,034)	0,605 (0,032)	0,458 (0,033)
LLDA	0,475 (0,029)	0,573 (0,023)	0,707 (0,026)	0,562 (0,021)	0,424 (0,034)	0,524 (0,025)	0,673 (0,026)	0,515 (0,022)
SEMCLA	0,559 (0,015)	0,616 (0,014)	0,719 (0,014)	0,61 (0,012)	0,499 (0,018)	0,559 (0,013)	0,68 (0,017)	0,552 (0,011)
Komitety klasyfikatorów, próba ucząca z Wikipedii L=2 S=200								
25x(B,W)	**0,599** (0,453)	**0,707** (0,427)	**0,79** (0,377)	**0,69** (0,432)	**0,529** (0,5)	**0,662** (0,473)	**0,742** (0,439)	**0,638** (0,481)
25xSEMCLA	0,554 (0,463)	0,617 (0,45)	0,698 (0,43)	0,611 (0,452)	0,493 (0,501)	0,559 (0,497)	0,654 (0,478)	0,553 (0,497)
25x(B, W, SEMCLA)	**0,626** (0,461)	**0,73** (0,417)	**0,824** (0,363)	**0,715** (0,427)	**0,588** (0,493)	**0,691** (0,463)	**0,797** (0,404)	**0,677** (0,468)

Tabela : Średnie wartości miar precyzji dla zbioru *DMOZ 15*. Parametr *L* oznacza „poziom” kategorii w grafie Wiki, *S* wielkość próby uczącej.

Dane z luką semantyczną

	Średnie z 25 przebiegów			Komitety		
Zbiór:	B	W	SC	25xSC	25x(B,W)	25x(B,W,SC)
Bankier	0,701 (0,04)	0,611 (0,051)	0,841 (0,012)	0,847 (0,36)	0,715 (0,452)	0,801 (0,4)
Forsal: Waluty	0,987 (0,006)	0,924 (0,032)	0,994 (0,001)	0,995 (0,073)	0,99 (0,098)	0,998 (0,046)
Forsal: Finanse	0,959 (0,012)	0,925 (0,022)	0,983 (0,003)	0,982 (0,133)	0,963 (0,189)	0,982 (0,133)
Ginekologia	0,617 (0,082)	0,581 (0,108)	0,801 (0,049)	0,813 (0,39)	0,723 (0,448)	0,809 (0,394)
Kardiologia	0,916 (0,025)	0,891 (0,046)	0,942 (0,009)	0,938 (0,241)	0,951 (0,217)	0,969 (0,174)
Onkologia	0,84 (0,026)	0,856 (0,04)	0,883 (0,016)	0,884 (0,32)	0,894 (0,308)	0,935 (0,247)

Tabela : Wartości średnie i odchylenia standardowe dla standardowej miary precyzji dla danych z luką semantyczną. Trzy pierwsze kolumny to średnie z 25 powtórzeń dla badanych zbiorów. Kolumny 4-5 to komitety klasyfikatorów.

Oznaczenia: B - Bayes, W - Balanced Winnow, SC - semantyczny klasyfikator

Przykład kategoryzacji dokumentów, taksonomia: Wikipedia

Przykład:
Kategoryzujemy treść dokumentu, jakim jest treść strony:
`http://www.nineplanets.pl/`:

> *Dziewięć Planet to przegląd historii, mitologii oraz aktualnej wiedzy naukowej związanej z wszystkimi planetami i księżycami naszego Układu Słonecznego. Każda z podstron zawiera tekst i zdjęcia, niektóre z nich także dźwięki i filmy, a większość zawiera także szereg odnośników do dalszych informacji. (...)*

Ranking trzech top kategorii wraz z wagami:

- układ słoneczny, waga: $0,137$
- planety, waga: $0,125$
- amerykańscy raperzy, waga: $0,089$.

Przykład kategoryzacji dokumentów, taksonomia: MeSH

Kategoryzujemy treść artykułu **The Activating Effect Of Magnesium And Other Cations On The Hemolytic Function Of Complement**, dysponując jego streszczeniem:

1. The evidence presented indicates that Mg, or other cation such as Ca, Ni, or Co, is essential for the hemolytic action of C'. Ca, Ni, and Co are less effective than Mg. The hemolytic system usually does not contain sufficient Mg for optimal hemolytic activity so that a marked enhancement can be obtained by addition of extra Mg. 2. The enhancing action of tissue fluids can be ascribed to their contribution of Mg. 3. Substances which bind Mg and Ca are anticomplementary when added to the usual hemolytic system which contains only a small quantity of Mg ...

Etykieta MESH nadana przez eksperta: [Complement System Proteins]

Ranking trzech top kategorii wraz z wagami:

- complement system proteins, waga: 0,141
- extravascular lung water, waga: 0,081
- health services research, waga: 0,059

Przykład działania algorytmu agregacji adaptacyjnej

Przykład:
Kategoryzujemy dokumenty ($N = 280$) pochodzące z `http://kopalniawiedzy.pl` z katalogu *astronomia/fizyka*.

Ranking trzech kategorii z największymi wagami przypisanych dokumentom:

- `łaziki` — waga: 12.0
- `astronautyka` — waga: 11.9
- `bozony` — waga: 10.0

Ranking trzech kategorii z największymi wagami przypisanych dokumentom gdy stosujemy **agregację adaptacyjną**:

- `astrofizyka` — waga: 78.8
- `galaktyki` — waga: 56.7
- `kosmologia` — waga: 16.7

Podsumowanie

W pracy zaproponowano:

- algorytm semantycznej kategoryzacji dokumentów,
- algorytm agregacji kategorii,
- rodzinę algorytmów semantycznych klasyfikatorów,
- heterogeniczny komitet klasyfikatorów (łączący algorytm semantycznej kategoryzacji i znanych dotąd klasyfikatorów).

Algorytm semantycznej kategoryzacji charakteryzuje się:

- działaniem bez konieczności posiadania specjalnie przygotowanego zbioru uczącego,
- pozwala zwracać wyniki z dowolnie określonego podzbioru taksonomii lub w sposób nienadzorowany rzutować wyniki kategoryzacji na podprzestrzeń właściwą dla zbioru dokumentów,
- dobrze działa dla wymagającego języka, jakim jest język polski,
- pozwala ominąć problem luki semantycznej,
- można go szybko aktualizować i dysponować algorytmem działającym na najbardziej aktualnych pojęciach.

Publikacje

- K. Ciesielski,P. Borkowski, M. A. Kłopotek, K. Trojanowski, K. Wysocki (2011) *Wikipedia-Based Document Categorization, Security and Intelligent Information Systems, SIIS, Warsaw, Poland*.
- P. Borkowski, K. Ciesielski *Etykietowanie dokumentów tekstowych z wykorzystaniem niejednorodnych komitetów klasyfikatorów i semantycznej kategoryzacji, Proceedings of Artificial Intelligence Studies. XIII International Conference on Artificial Intelligence AI-26'2012*
- P. Borkowski, K. Ciesielski, M. A. Kłopotek (2014) *Unsupervised Aggregation of Categories for Document Labelling, Foundations of Intelligent Systems - 21st International Symposium, ISMIS 2014.*
- P. Borkowski, K. Ciesielski, M. A. Kłopotek (2017) *The role of semantic similarity in document classification, Arxiv*