STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Julian SZYMAŃSKI, Marcin DEPTUŁA, Henryk KRAWCZYK
Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki,
Katedra Architektury Systemów Komputerowych

IDENTYFIKACJA POWIĄZAŃ POMIĘDZY KATEGORIAMI WIKIPEDII Z UŻYCIEM MIAR PODOBIEŃSTWA ARTYKUŁÓW

Streszczenie. W artykule opisano podejście do identyfikacji powiązań między kategoriami w repozytorium danych tekstowych, bazując na Wikipedii. Przeprowadzając analizę podobieństwa między artykułami, określono miary pozwalające zidentyfikować powiązania między kategoriami, które nie były wcześniej uwzględnione, i nadawać im wagi określające stopień istotności. Przeprowadzono automatyczną ocenę uzyskanych rezultatów w odniesieniu do już istniejącej struktury kategorii.

Słowa kluczowe: kategoryzacja tekstu, identyfikacja asocjacji, Wikipedia

IDENTIFICATION OF WIKIPEDIA CATEGORIES ASSOCIATIONS BASED ON ARTICLES SIMILARITIES

Summary. In the article we present an approach to identification of relations between categories organizing the repository of documents. We describe the metrics of category relevance based on similarity measures between articles. The metrics have been used to discover relations between categories within Wikipedia repository. The evaluation of the proposed method indicate it allows to reconstruct already existing associations in category structure as well as introduce new significant relations.

Keywords: text categorization, association recognition, Wikipedia

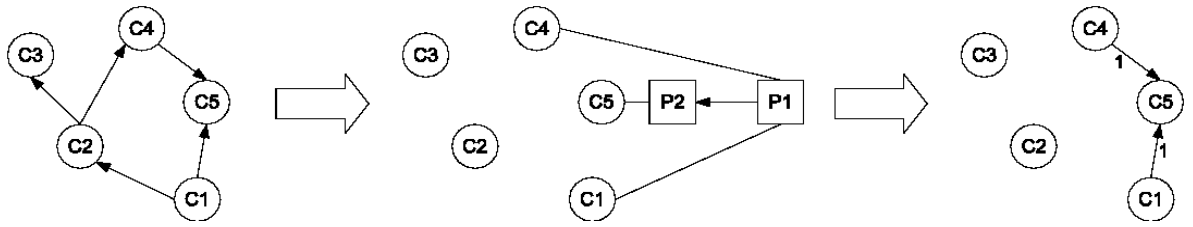
1. Wstęp

Wikipedia jest obszernym zbiorem wiedzy dostępnym on-line. Jego zawartość jest tworzona przez wolontariuszy, w związku z tym jest ogólnodostępna do wglądu i edycji. Ten mechanizm kooperacyjnej pracy powoduje, że dane wprowadzane do Wikipedii nie są poddawane rygorystycznej kontroli, co może stwarzać problemy związane z jakością informacji

w niej zawartych. Otwarcie na powszechną edycję pozwoliło na bardzo szybki wzrost treści encyklopedii. Tak duży zbiór powiązanych ze sobą informacji tekstowych stwarza szerokie możliwości analizy danych językowych, dlatego z punktu widzenia repozytoriów języka naturalnego Wikipedia jest interesującym zasobem badawczym.

W artykule przeprowadzono analizę struktury powiązań pomiędzy artykułami w celu identyfikacji powiązań między kategoriami organizującymi wiedzę w Wikipedii. Podstawowym założeniem tej analizy jest to, że zależności między kategoriami można określić na podstawie miary podobieństwa między artykułami do nich należącymi. Przyjęto, że przypisanie artykułów do kategorii, a także zależności pomiędzy artykułami są dane a priori. Zadaniem jest tu stworzenie miary pozwalającej na określenie wartości podobieństwa pomiędzy artykułami i określenie na tej podstawie miary podobieństwa kategorii.

Proces ten zobrazowano na rys. 1. Do oznaczeń przyjęto symbole *C* i *P* odpowiednio dla kategorii i artykułów. Sytuacją wejściową jest zadany graf kategorii (z lewej). Następnie zostają usunięte krawędzie z tego grafu opisujące połączenia pomiędzy kategoriami i zostają dołączone informacje o przynależności artykułów do kategorii (rysunek środkowy). Na podstawie analizy podobieństwa pomiędzy artykułami należącymi do poszczególnych kategorii zostaje wyznaczony nowy, ważony graf kategorii (przedstawiony na rysunku z prawej) [1, 2]. Tak uzyskany nowy graf kategorii można porównać z dotychczas istniejącym, a także ocenić pod kątem kompletności i poprawności. Przeprowadzono to w punktach przedstawiających rezultaty badań.



Rys. 1. Wizualizacja procesu generowania powiązań pomiędzy kategoriami na podstawie podobieństwa pomiędzy artykułami
Fig. 1. Graphical presentation of the categories associations identification process

2. Pozyskiwanie danych

Do pozyskania danych zostały wykorzystane pliki udostępnione przez Wikimedia Foundation w formie spakowanych plików w formacie SQL. W eksperymentach została użyta wersja SimpleWiki z dnia 01.04.2010 r. Dane zawężono do superkategorii *Category:Articles*, czyli kategorii, które są nadrzędne wobec wszystkich innych kategorii w bazie danych. Pozostałe superkategorie, takie jak: kategorie zalążkowe (ang. *stub*), np. *Category:People_stubs*, lub kategorie grupujące artykuły wymagające uporządkowania, np. *Category:Cleanup-*

_needed, zawierają informacje dodatkowe i wspierające prace redaktorów. Z punktu widzenia przeprowadzanej analizy nie wnoszą one nic wartościowego, a w związku z niewielką zawartością treści mogą wpływać niekorzystnie na jakość uzyskanych wyników.

Do ekstrakcji danych została wykorzystana aplikacja Matrixu [3] umożliwiająca zamianę Wikipedii na postać maszynowo przetwarzaną. Dzięki temu otrzymano:

- słownik artykułów (tabela 1 a),
- słownik kategorii (tabela 1 b),
- powiązania między artykułami (tabela 1 c),
- powiązania między artykułami a kategoriami (tabela 1 d),
- grafy kategorii (tabela 1 e).
- listy artykułów przynależących do poszczególnych kategorii (tabela 1 f),

Tabela 1

Format danych wejściowych

a) słownik artykułów	article1_id\t**article1_name\r\n** article2_id\t**article2_name\r\n** ...
b) słownik kategorii	category1_id\t**category1_name\r\n** category2_id\t**category2_name\r\n** ...
c) powiązania między artykułami	article1_id#assoc1_article_id-num_of_links1 assoc2_article_id-num_of_links2\n article2_id#assoc3_article_id-num_of_links3\n ...
d) powiązania między artykułami, a kategoriami	article1_id\tcategory1_id\t**category2_id\n** article2_id\t**category3_id\n** ...
e) drzewo kategorii	category1_page_id\t**assoc1_category_page_id** assoc2_category_page_id\r\n category2_page_id\t**assoc3_category_page_id\r\n** ...
f) lista artykułów	category1_page_id\t**article2_id\r\n** category3_page_id\t**article4_id\r\n** ...

W tabeli 1 wykorzystano następujące oznaczenia:

article_id – identyfikator artykułu,

article_name – tytuł artykułu,

category_id – identyfikator kategorii,

category_name – nazwa kategorii,

assoc_article_id – identyfikator artykułu powiązanego z artykułem wymienionym na początku wiersza (article_id),

num_of_links – liczba powiązań między artykułem źródłowym (article_id) a artykułem powiązanym (assoc_article_id),
category_page_id – identyfikator strony opisującej kategorię,
assoc_category_page_id – identyfikator strony opisującej kategorię powiązaną z kategorią, do której należy strona wymieniona na początku linii (category_page_id).

SimpleWiki jest jedną z wielu istniejących Wikipedii. Przykładowe rozmiary Wikipedii w różnych wersjach językowych przedstawiono w tabeli 2. To, co wyróżnia SimpleWiki spośród innych Wikipedii, to fakt, że nie jest ona tworzona w żadnym z języków narodowych, tylko w swego rodzaju współczesnym esperanto, którym jest Simple English. Język ten składa się wyłącznie z podstawowego słownictwa języka angielskiego, które w artykułach jest przeważnie wykorzystywane w postaci krótkich zdań o prostej gramatyce. Ma to na celu umożliwienie tworzenia zbioru podstawowej wiedzy opisanej prostym językiem, przystępnym dla osób, które dopiero uczą się języka angielskiego. Dodatkowo jedną z wytycznych dla twórców artykułów w SimpleWiki jest ograniczenie długości pojedynczego artykułu do maksymalnie 1000 słów.

Tabela 2

Porównanie rozmiarów przykładowych Wikipedii

	Liczba kategorii	Liczba artykułów	Liczba połączeń pomiędzy kategoriami	Liczba połączeń pomiędzy artykułami	Średnia liczba linków do kategorii zależnych w kategorii	Średnia liczba linków w artykule	Gęstość grafu kategorii	Gęstość grafu artykułów
1) EnWiki	797 291	18 021 542	37 154 690	405 023 273	46,60	22,47	5,84E-05	1,25E-06
2) DeWiki	165 798	3 133 254	4 703 753	62 946 323	28,37	20,09	1,71E-04	6,41E-06
3) PlWiki	111 670	1 416 520	2 175 515	43 656 852	19,48	30,82	1,74E-04	2,18E-05
4) SimpleWiki	25 072	155 101	204 499	3 548 863	8,16	22,88	3,25E-04	1,48E-04

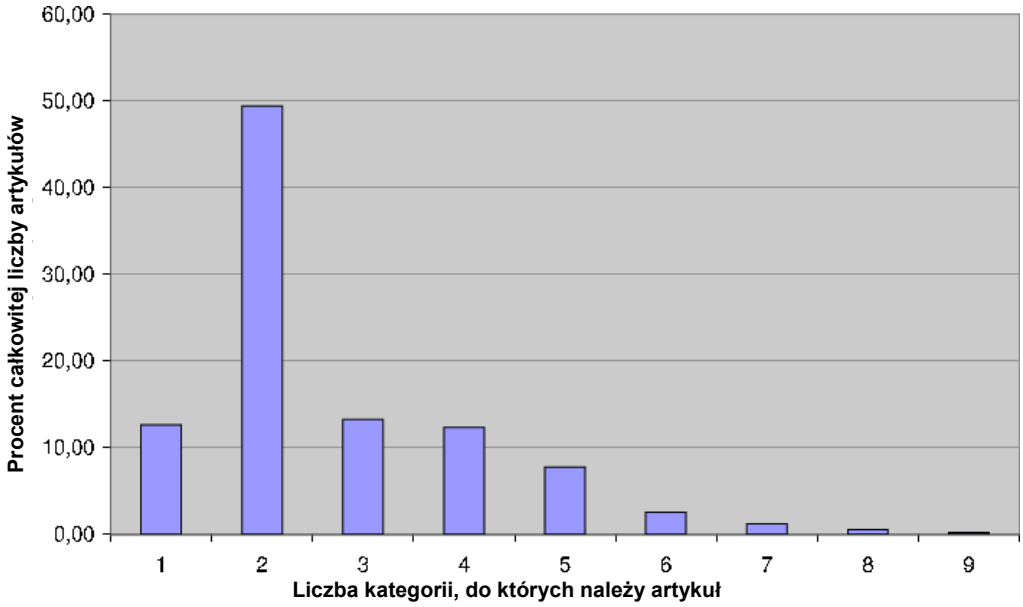
Cechy te, specyficzne dla SimpleWiki, powodują, że SimpleWiki jest uproszczoną wersją EnWiki. Dzięki temu SimpleWiki jest dobrym poligonem doświadczalnym do badań nad EnWiki, której rozmiar powoduje konieczność uwzględniania dodatkowych kwestii związanych z wydajnością przetwarzania dużych ilości danych.

3. Ogólne statystyki danych wejściowych

Ze wstępnej analizy uzyskanych danych zostały wyciągnięte podstawowe statystyki, które są istotne dla oceny wyników uzyskanych w eksperymentach.

Dane wejściowe zawierały 155 101 artykułów zgrupowanych w 25 072 kategoriach. 87% artykułów należało do więcej niż jednej kategorii. Szczegółowy rozkład udziału artykułów w funkcji liczby kategorii, do której przynależą, przedstawiono na rys. 2.

Artykuły były powiązane 3 548 863 linkami. Daje to średnio 22,88 linków (powiązań z użyciem hiperreferencji) wychodzących z artykułu. Rozkład liczby linków na poszczególne artykuły nie jest równomierny. Wyznaczenie średniej liczby linków i wykreślenie jej w funkcji długości artykułu (rys. 3) pokazało, że liczba linków rośnie wraz z rozmiarem artykułu – co jest zgodne z oczekiwaniami. Współczynnik kierunkowy linii trendu równy 0,0009969 wskazuje, że średnio na każde kolejne 1000 znaków treści artykułu przypada jeden link więcej.



Rys. 2. Procentowy udział artykułów w funkcji liczby kategorii, do których należą
Fig. 2. The distribution of the articles amount and number of their categories

Ze względu na duży krok osi X i dużą ziarnistość danych średnia liczba linków przedstawiona na rys. 3 została wyliczona wg wzoru:

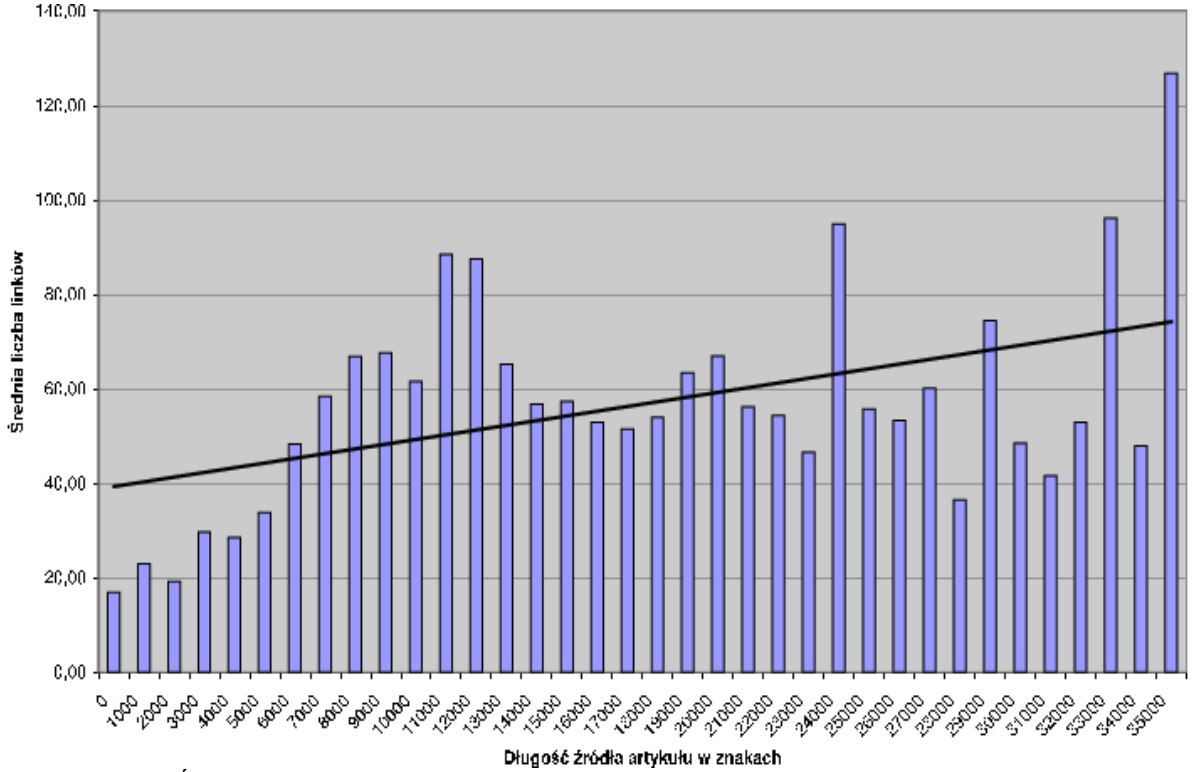
$$L_{sr(g_d,g_g)} = \sum_{|A_i| \in \langle g_d, g_g \rangle} \frac{|L(A_i)| * g_g}{|A_i|}, \tag{1}$$

gdzie: $L_{sr}(g_d, g_g)$ – średnia liczona w granicach $\langle g_d, g_g \rangle$, $|A_i|$ – długość artykułu A_i, $|L(A_i)|$ – liczba linków wychodzących z artykułu A_i.

4. Cel analizy

Zawartość Wikipedii składa się z kategorii i artykułów, które się w nich znajdują. Kategorie są pomiędzy sobą powiązane, a treść artykułów zawiera odwołania do innych artykułów związanych tematycznie z danym zagadnieniem.

Zarówno kategorie, artykuły, powiązania pomiędzy artykułami, jak i związki pomiędzy kategoriami są tworzone ręcznie przez redaktorów – wolontariuszy pracujących dla Wikipedii. Celem przeprowadzonych badań jest analiza możliwości automatycznego identyfikowania nowych powiązań pomiędzy kategoriami opisującymi treść artykułów w nich zawartych na wysokim poziomie abstrakcji. Pozwoli to na wynajdowanie nowych istotnych powiązań pomiędzy kategoriami, które nie zostały jawnie zdefiniowane przez redaktorów, a mogą być interesujące dla osób korzystających z opisu kategorialnego treści Wikipedii. Podejście takie będzie umożliwiać np. znajdowanie nowych interesujących kierunków, w których można przeprowadzić wyszukiwanie informacji [4].

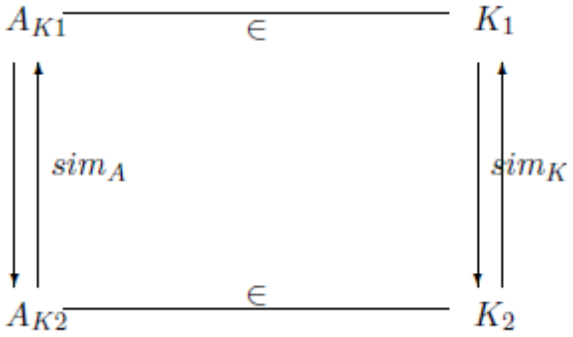


Rys. 3. Średnia liczba linków w funkcji długości artykułu w znakach wraz z linią trendu
Fig. 3. The average links number in function of article length (in characters)

Ogólniej problem sprowadza się do utworzenia miary podobieństwa pomiędzy kategoriami sim_K zgodnie ze wzorem (2), na podstawie wybranej miary podobieństwa pomiędzy artykułami sim_A.

$$sim_K(K_i, K_j) \approx sim_A(A \in K_i, A \in K_j) \tag{2}$$

Oznaczenie ≈ określa utworzenie podobieństwa pomiędzy kategoriami K_i oraz K_j na podstawie podobieństwa sim_A – artykułów należących do każdej z kategorii $\left(A \in K_i, A \in K_j\right)$. Proces ten zobrazowano na rys. 4.



Rys. 4. Analiza podobieństwa kategorii na podstawie podobieństwa pomiędzy artykułami w nich zawartymi
Fig. 4. Analysis of categories similarity based on their articles similarity

5. Funkcja podobieństwa pomiędzy artykułami

Funkcja sim_A może zostać zdefiniowana różnorodnie, zależnie od tego, jakie cechy artykułów są uwzględnione w mierzonym podobieństwie. Dla celów tej analizy przyjęto miarę opartą na referencjach występujących pomiędzy artykułami. Intencją tej miary jest określenie, że dwa artykuły są tym bardziej podobne, im silniej są ze sobą powiązane linkami.

Przykładami innych miar mogłoby być chociażby występowanie podobnych słów w treści bądź określanie podobieństwa w analizie spakowanych plików danych [5]. Założono, że analiza zostaje wykonana wyłącznie dla linków bezpośrednich, jednakże można rozważyć powiązania wyższych rzędów, ze względu na strukturę małego świata [6, 7] Wikipedii, które muszą zostać w odpowiedni sposób ważone. Inną możliwością jest użycie informacji o kolejności pojawiania się referencji w artykule i np. wykorzystywanie tylko tych, które znajdują się na początku artykułu (ze względu na ich większą deskryptywność) lub też zastosowanie metod rankingowania [8, 9], co pozwoliłoby określić relewantność linków.

Przyjęto, że artykuły A_i i A_j są do siebie tym bardziej podobne, im większy jest udział liczby linków bezpośrednio je łączących $\left|L\left(A_i, A_j\right)\right|$ w sumarycznej liczbie linków zdefiniowanych w tej parze artykułów $\left|L\left(A_i\right)\right| + \left|L\left(A_j\right)\right|$.

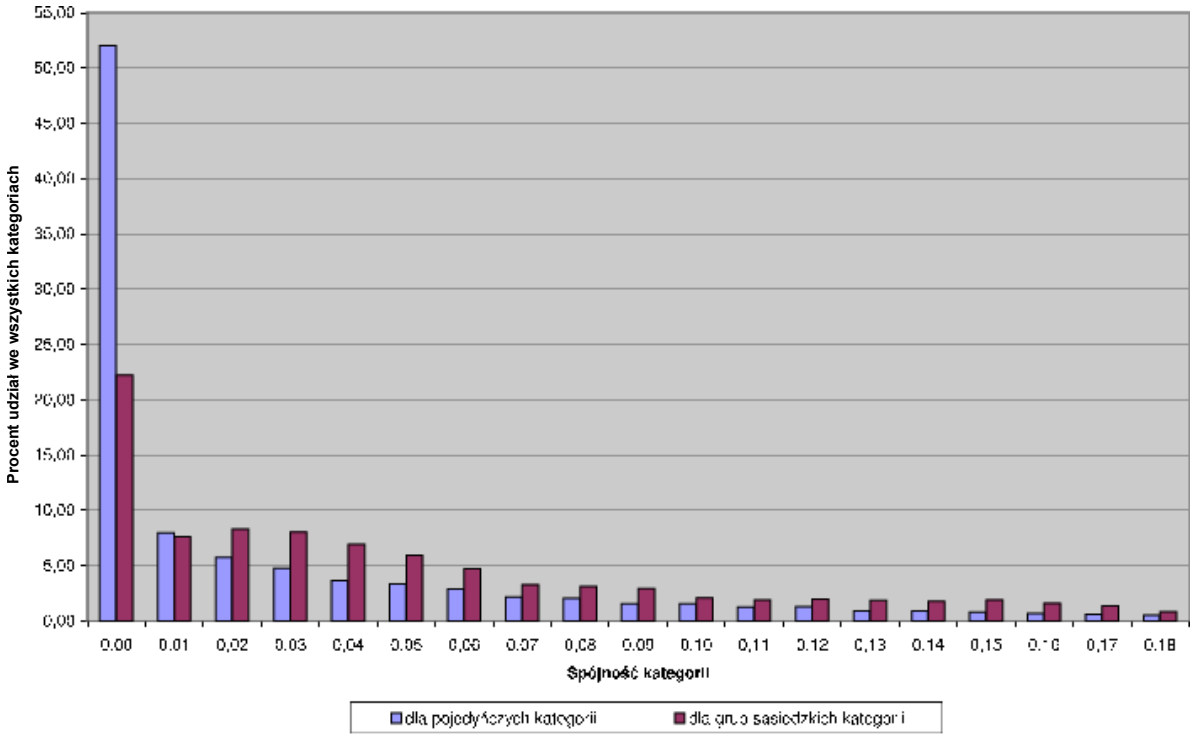
$$sim_A\left(A_i, A_j\right) = \frac{\left|L\left(A_i, A_j\right)\right|}{\left|L\left(A_i\right)\right| + \left|L\left(A_j\right)\right|}, \tag{3}$$

gdzie: A_j – artykuł, $L\left(A_i, A_j\right)$ – zbiór linków łączących artykuły A_i i A_j, $L(A_i)$ – zbiór wszystkich linków wychodzących z A_i.

Dla tak zdefiniowanej funkcji podobieństwa wyznaczono statystykę podobieństwa dla zbioru wszystkich par artykułów. Statystyka ta pokazuje, że dowolnie wybrane dwa artykuły w 99,9994% przypadków nie przekroczą 0,1 podobieństwa. Ze szczegółowych danych wynika, że podobieństwo dla losowo wybranej pary nie przekroczy nawet 0,01 w 99,95% przypadków. Wskazuje to na to, że przyjęta miara podobieństwa nie powoduje dość znaczących korelacji dla losowo wybranych artykułów, co jest zgodne z oczekiwaniami. Jednocześnie można z tego wywnioskować, że graf podobieństwa pomiędzy artykułami wygenerowany przy użyciu tej miary jest bardzo rzadki. Potwierdzają to statystyki przedstawione w tabeli 2.

Oceniając jakość takiego grafu, należy zbadać, czy podobieństwo pomiędzy artykułami jest większe tam, gdzie się tego spodziewano. Takie wyniki oczekiwane są dla artykułów należących do jednej kategorii. Analiza tego przypadku została wykonana w ramach oceny spójności kategorii opisanej w punkcie 6.

6. Spójności kategorii



Rys. 5. Procentowy udział kategorii w funkcji ich spójności
Fig. 5. Distribution of categories cardinality in function of their coherence

Na podstawie zdefiniowanej w punkcie 5 funkcji podobieństwa sim_A zdefiniowano funkcję spójności kategorii $S(K_i)$ jako proporcję liczby linków łączących wyłącznie artykuły w obrębie pojedynczej kategorii do liczby wszystkich linków zdefiniowanych w artykułach do niej należących. Można to wyrazić przez sumę podobieństw pomiędzy poszczególnymi

parami artykułów ważoną liczbą linków wychodzących z artykułów pary o danym podobieństwie i normalizowaną przez liczbę wszystkich linków w obrębie kategorii.

$$S(K_i) = \frac{\sum_{A_j, A_k \in K_i} \left(sim_A(A_j, A_k) * \left(|L(A_j)| + |L(A_k)| \right) \right)}{\sum_{A_j \in K_i} |L(A_j)|} = \frac{\sum_{A_j, A_k \in K_i} L(A_j, A_k)}{\sum_{A_j \in K_i} |L(A_j)|} \quad (4)$$

Oznaczenia przyjęte we wzorze (4) są takie same jak we wzorze (3). Dodatkowym założeniem jest to, że kategorie (K_i) są zbiorami zawierającymi artykuły, które są do nich przypisane.

Dla tak zdefiniowanej miary spójności wyliczono jej wartości dla wszystkich kategorii i ujęto je w statystyce na rys. 5. Spójność okazała się równa 0,000 dla 52% kategorii wg założonej miary spójności. Dla pozostałych 48% wartość ta okazała się niezerowa. Średnia spójność w obrębie tych 48% kategorii jest ponaddwukrotnie większa niż średnia spójność wszystkich kategorii i równa niemal 0,1. Świadczy to o tym, że dla prawie połowy kategorii miara oparta na linkach ma zastosowanie i wykazuje zależność artykułów w obrębie kategorii. Druga połowa kategorii odstaje pod tym względem, co bezpośrednio świadczy o małej liczbie powiązań pomiędzy artykułami należącymi do kategorii z tej grupy.

Zakładając, że artykuły z kategorii sąsiednich (należących do jednego rodzica) mogą być powiązane większą liczbą referencji niż te w obrębie pojedynczej kategorii, zgrupowano kategorie w grupy sąsiedzkie. Statystyka spójności dla tak zdefiniowanych zbiorów okazała się lepsza niż dla pojedynczych kategorii. Obie statystyki zostały porównane na rys. 5. W statystyce opartej na grupach sąsiedzkich udział jednostek o spójności w przedziale (0,00-0,01) spadł o połowę. Średnia spójność, która dla pojedynczych kategorii była równa tyko 0,04, dla grup kategorii okazała się dwukrotnie wyższa. Świadczy to o tym, że co najmniej drugie tyle powiązań, które łączą artykuły w obrębie samej kategorii, wiąże artykuły z sąsiednich, spokrewnionych kategorii. Z tego powodu średnia spójność dla grup sąsiedzkich jest odpowiednio wyższa.

Wyodrębniono 1 139 kategorii, które wykazują spójność poniżej 0,01, biorąc je pod uwagę zarówno jako pojedyncze kategorie, jak i w ramach grupy sąsiedzkiej. Kategorie te stanowią 4,5% wszystkich kategorii.

Analizując ten zbiór okazało się, że 666 (2,7% wszystkich kategorii) z nich nie zawiera artykułów. Są to takie kategorie, jak np. *Category:Content*, w których wiele artykułów się zawiera, jednak ze względu na reguły przypisywania artykułów do kategorii artykuły te są przypisane do ich kategorii potomnych[1].

Pozostałe 1,8% kategorii, które uzyskały spójność poniżej 0,01 zarówno pojedynczo, jak i w grupie sąsiedzkiej, stanowi bardzo interesującą grupę, która zgodnie z założeniami przy-

[1] http://en.wikipedia.org/wiki/Wikipedia:Categorization

jętej miary faktycznie wykazuje zerową spójność. Należy do niej np. *Category:Greek_people*, zawierająca w sobie artykuły: *Dave_Batista*, *El_Greco*, *Prince_Philip*, *Duke_of_Edinburgh* i *Mike_Zambidis*, które rzeczywiście nie mają ze sobą nic wspólnego poza tym, że opisują osoby, pochodzące z Grecji, jednakże nie są zupełnie powiązane.

Innym przykładem takiej kategorii jest *Category:Natural_sciences*, zawierającą tylko artykuł *Biology*, który, ponieważ nie ma powiązań sam ze siebą, nie zapewnia kategorii spójności. Dodatkowo ze względu na to, że ta kategoria ma dość wysoki poziom abstrakcji, jej grupa sąsiedzka (*Category:Biology, Category:Physical_sciences*) także dotyczy pojęć na tyle abstrakcyjnych, że nie istnieje żaden link łączący artykuły należące do tych kategorii.

7. Ocena nowych powiązań pomiędzy kategoriami w odniesieniu do powiązań zdefiniowanych przez redaktorów Wikipedii

Funkcje podobieństwa kategorii sim_K zdefiniowano jako liczbę linków łączących artykuły należące do tych kategorii w stosunku do całkowitej liczby linków w nich zdefiniowanych. Można to wyrazić jako znormalizowaną, ważoną sumę podobieństw par artykułów z tych kategorii.

$$\begin{aligned} sim_K\left(K_i, K_j\right) &= \frac{\sum_{A_k \in K_i} \sum_{A_l \in K_j} \left(sim_A\left(A_l, A_k\right) * \left(\left|L\left(A_l\right)\right| + \left|L\left(A_k\right)\right|\right)\right)}{\sum_{A_k \in K_i} \left|L\left(A_k\right)\right| + \sum_{A_l \in K_j} \left|L\left(A_l\right)\right|} \\ &= \frac{\sum_{A_k, A_l \in K_i} L\left(A_k, A_l\right)}{\sum_{A_k \in K_i} \left|L\left(A_l\right)\right|} \end{aligned} \tag{5}$$

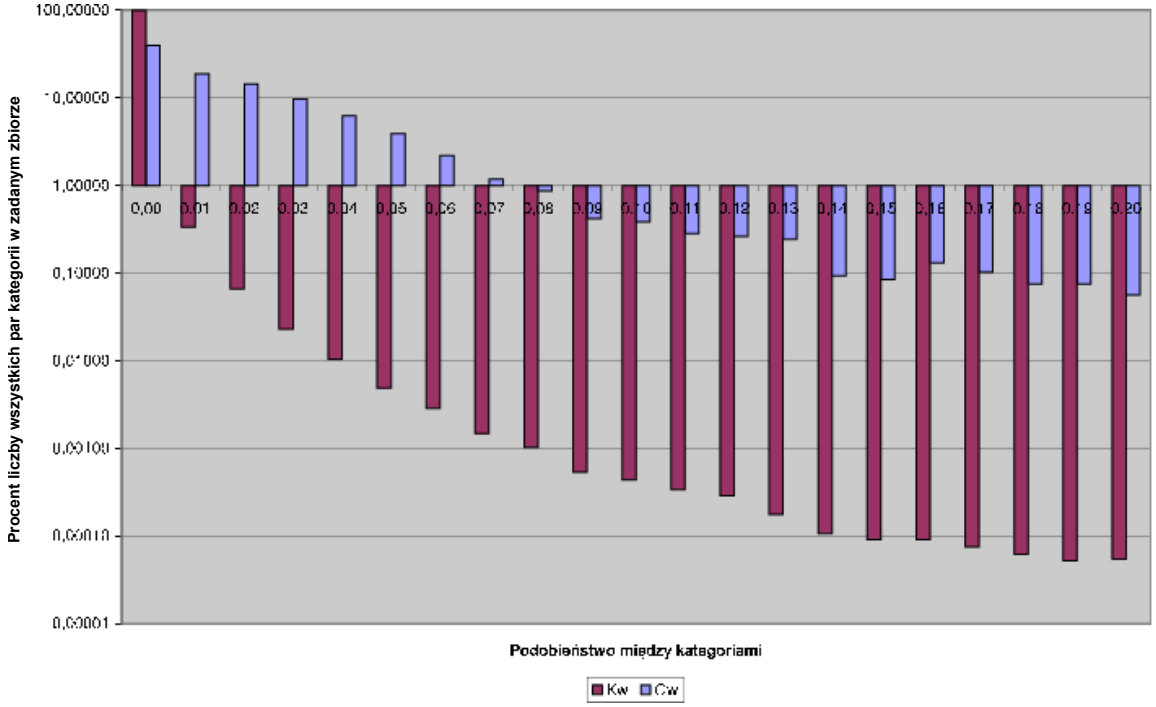
We wzorze (5) przyjęto oznaczania jak we wzorach (3) i (4). Dla tak zdefiniowanej miary podobieństwa została wyznaczona wartość podobieństwa dla wszystkich par kategorii. Niemalże wszystkie pary kategorii wykazały zerowy poziom podobieństwa. Dla dowolnie wybranej pary kategorii istnieje 99,55% prawdopodobieństwa, że te kategorie nie będą ze sobą w ogóle związane. Jest to uzasadnione, gdyż prawdopodobieństwo wylosowania pary pokrewnych kategorii ze zbioru 38 592 505 możliwych par kategorii, które zostały przyjęte do analizy, jest bardzo małe.

Po przeanalizowaniu grafu kategorii zdefiniowanego w wersji SimpleWiki (G_w) okazało się, że składa się on z 10 668 połączeń definiujących pary kategorii, które są ze sobą bezpośrednio związane. Stanowi to 0,028% wszystkich możliwych par kategorii, co wskazuje, że wynik statystyki podobieństwa kategorii uzyskany dla dowolnych par kategorii jest uzasadniony.

Miara podobieństwa (wzór (3)) zastosowana dla grafu G_w już tylko dla niecałych 40% dała wartości poniżej 0,01. Dla pozostałych 60% par wykazała znaczące podobieństwo. Jest to

diametralnie inna statystyka niż dla zbioru dowolnych par kategorii (K_w). Porównanie obu statystyk zostało przedstawione na rys. 6.

Szukając odpowiedzi na pytanie, dlaczego mimo wszystko 40% par kategorii wykazało tak niską wartość podobieństwa z użyciem funkcji opartej na linkach (wzór (5)), można zwrócić uwagę na to, że bardzo podobny procent kategorii wykazywał brak spójności (50%). Wobec powyższego odrzuciliśmy wszystkie pary zawierające kategorie o wartości spójności poniżej 0,01 (powstały graf został oznaczony jako G_{ws}). Spowodowało to odrzucenie 6947 par kategorii (~65%). Pozostałe 3721 par okazało się mieć już znacznie lepszą statystykę. W tej grupie już tylko trochę ponad 25% par kategorii wykazało podobieństwo w przedziale (0,00 – 0,01) i niemalże drugie tyle znalazło się w przedziale (0,01 – 0,02). Procentowy udział par kategorii maleje też wyraźnie wolniej w kierunku rosnących wartości podobieństwa niż dla grafu (G_w) sprzed odrzucenia par zawierających kategorie o małej wartości spójności.

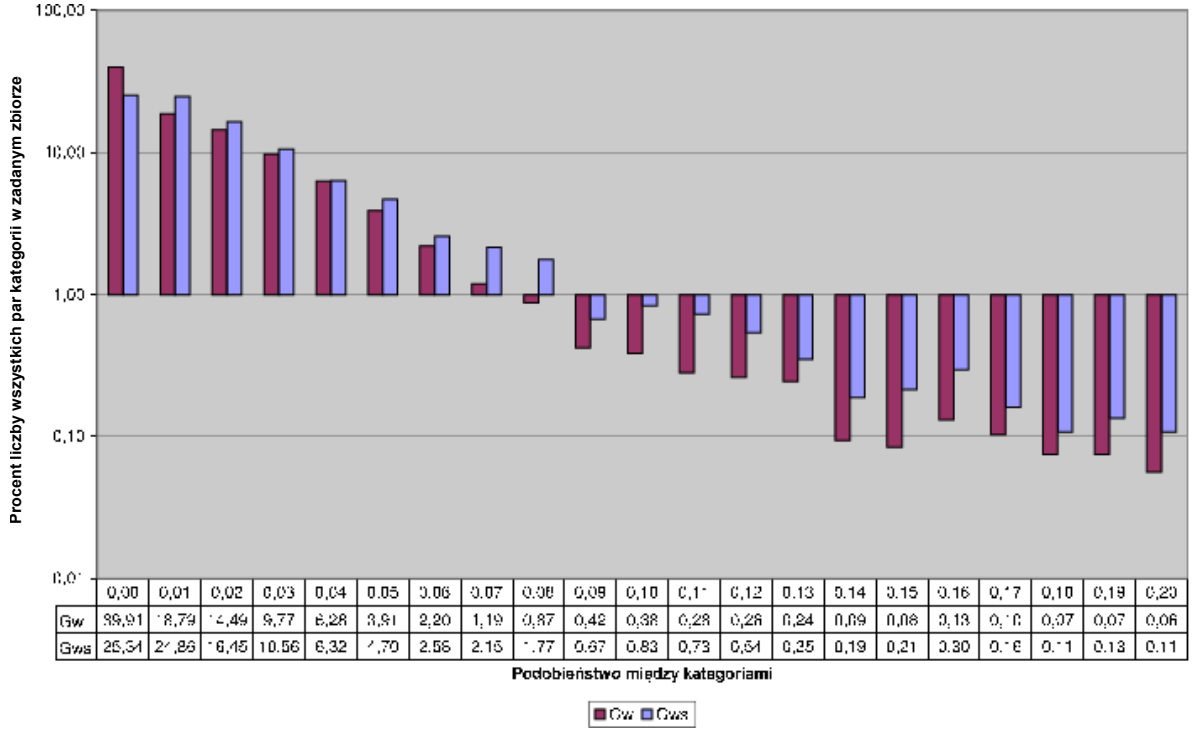


Rys. 6. Porównanie statystyki podobieństwa wyznaczonej dla par należących do grafu kategorii (G_w) i dla wszystkich możliwych par kategorii (K_w)

Fig. 6. Comparison of similarity distribution calculated for pairs taken form graph category (G_w) and all possible pairs of categories (K_w)

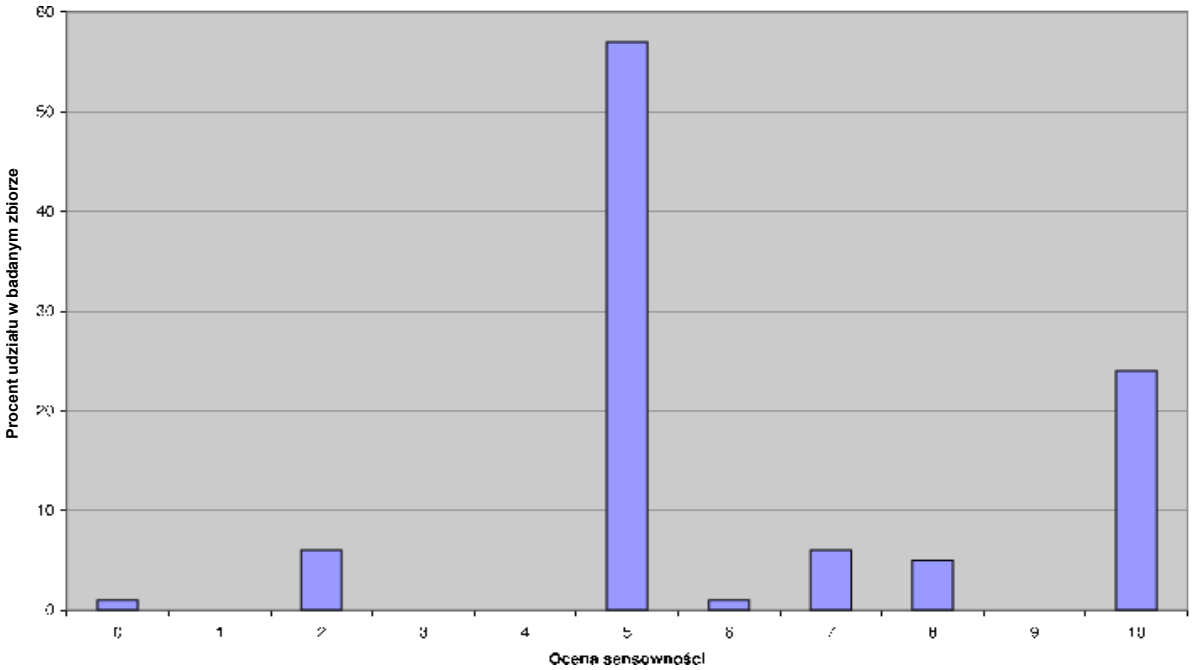
Podczas analizy grafu G_{ws} wyodrębniono 954 pary kategorii, które wykazały spójność poniżej 0,01. Pary te stanowią 0,9% wszystkich połączeń pomiędzy kategoriami. Z bezpośredniej analizy tych połączeń wynikło, że pomimo iż pomiędzy kategoriami utworzone zostały słabe powiązania, to brak powiązań artykułów między sobą jest uzasadniony. Przykładem takiego połączenia może być połączenie pomiędzy kategoriami *Category:Water* i *Category:Water_transportation*. Kategoria *Category:Water_transportation* zawiera się w kategorii *Category:Water*, ale ponieważ pierwsza zawiera takie artykuły, jak np.: *Aqueduct*, *Hydro-*

logy, *Vapor*, a kategoria *Category:Water_transportation* zawiera takie artykuły, jak: *Crew*, *Dock* i *Ferry*, to uzasadnione jest, że te dwie grupy artykułów się ze sobą nie łączą.



Rys. 7. Porównanie statystyki podobieństwa wyliczonej dla par należących do grafu kategorii (G_w) i do zmodyfikowanego grafu kategorii (G_{ws})

Fig. 7. Comparison of similarity distribution calculated for pairs of category graph (G_w) and modified category graph (G_{ws})



Rys. 8. Statystyka sensowności połączenia dla najbardziej podobnych par kategorii spoza grafu G_w

Fig. 8. Distribution of meaningful associations between the most similar categories out of G_w graph

8. Ocena nowych połączeń pomiędzy kategoriami, niezdefiniowanych wcześniej w grafie kategorii Wikipedii

Korzystając z przyjętej miary podobieństwa pomiędzy kategoriami, można postawić pytanie: na ile graf kategorii zdefiniowany w SimpleWiki jest kompletny? W celu dokonania tej oceny pobrano 1000 par kategorii spoza grafu kategorii o możliwie największej wartości podobieństwa. Otrzymany zbiór par reprezentował podobieństwa w przedziale (0,17 – 0,62), czyli większe niż ponad 99% par z grafu G_w! Średnia wartość podobieństwa dla tego zbioru wyniosła 0,24.

W celu oceny jakości grafu powstałego po dodaniu nowo otrzymanych połączeń do oryginalnego grafu kategorii przeprowadzono ręczną ocenę sensowności. Ocena ta była przyznawana w skali 0-10 i została przeprowadzona niezależnie przez dwie osoby. Otrzymane uśrednione wyniki przedstawiono na rys. 8. Obrazują one, że znaczna większość (93%) par kategorii z zadanego zbioru okazała się sensowna, otrzymując ocenę 5 lub większą. Jednak tylko ¼ z nich (24%) okazała się na tyle sensowna, żeby mieć pewność co do tego, że takie połączenie mogłoby istnieć i uzyskać ocenę 10. Znaczna większość, bo 57% ze wszystkich ocenianych par, otrzymała ocenę 5. Świadczy to o dużej zbieżności tematyki, jednak również o braku możliwości zdefiniowania sensownej relacji pomiędzy nimi. Na przykład *Category:Nirvana_albums* i *Category:Nirvana_songs* dotyczą dokładnie tej samej dziedziny, ale nie można jednoznacznie powiedzieć, że jedna zawiera drugą, bo każda z nich reprezentuje inny porządek tego samego zagadnienia. Ciekawą grupę stanowią pary, które otrzymały oceny z zakresu 6-8. Są to kategorie, które się wzajemnie zawierają, ale nie jest to zawieranie pełne, jak np. *Category:Early_Christian_Saints* i *Category:New_Testament_people*. Wikipedia niestety nie przewiduje możliwości określenia relacji częściowego zawierania się kategorii. Relacje takie można jednak zbudować zarówno na podstawie istniejących połączeń pomiędzy kategoriami i artykułami, jak i na podstawie analizy ich treści czy struktury, przyjmując odpowiednie miary podobieństwa. Tworzenie takich połączeń oraz możliwości ich wykorzystania zostały szerzej opisane w [10]. Jak wynika z przeprowadzonych analiz, powiązania takie mogłoby być również interesujące z punktu widzenia użytkownika Wikipedii szukającego treści podobnych do zadanych. Takie powiązania pozwalałyby na przemieszczanie się po kategoriach w kierunku horyzontalnym, co stanowi rozszerzenie typowego poruszania się po strukturze hierarchicznej. Ten typ podobieństwa pozwala na identyfikację zagadnień związanych częściowo, a przez to pokrewnych koncepcyjnie.

9. Podsumowanie

Z analizy rezultatów uzyskanych w przeprowadzonych eksperymentach wynika, że miara podobieństwa oparta na powiązaniach jest miarą dającą dobre rezultaty. Na zadanych zbiorach danych otrzymano rezultaty zgodne z oczekiwaniami. Jednak słabym punktem tej miary z całą pewnością są dane wejściowe, bo w momencie gdy możliwe do zdefiniowania powiązania nie zostały utworzone w obrębie artykułów, wartość zarówno spójności, jak i podobieństwa spada praktycznie do zera. Taki efekt można było zaobserwować dla około 50% kategorii, co w sposób wyraźny rzutowało na wyniki.

Rozwiązaniem problemów miary opartej na powiązaniach między artykułami jest wyłącznie naprawa danych wejściowych. Potencjalnie nie będą miały tego problemu miary oparte na analizie treści artykułów, co planuje się zbadać w kolejnych częściach eksperymentu.

Uzyskane powiązania pomiędzy kategoriami zostały wykorzystane w naszym prototypowym systemie do wspomagania wyszukiwania informacji w Wikipedii (http://kask.eti.pg-.gda.pl/BetterSearch). W systemie tym powiązania pomiędzy kategoriami umożliwiają odnajdowanie informacji koncepcyjnie podobnych do tych, które wskazał użytkownik [11].

Zaproponowane podejście do identyfikacji powiązań między kategoriami na podstawie analizy podobieństwa obiektów może zostać uogólnione na innego rodzaju zbiory danych niż przedstawione w artykule repozytoria tekstu. Na przykład opisana metoda identyfikacji powiązań między elementami grup może zostać wykorzystana do odtwarzania asocjacji między grupami organizującymi podobne osoby w sieciach społecznościowych [12].

BIBLIOGRAFIA

1. Szymański J.: Mining relations between Wikipedia Categories. Proceedings of the 2th International conference of network of Digital Technologies, Springer, Prague 2010.
2. Holloway T., Bozicevic M., Borner K.: Analyzing and visualizing the semantic coverage of Wikipedia and its authors. Complexity, No. 12(3), 2007, s. 30÷40.
3. Szymański J.: Wikipedia Articles Representation with Matrix'u. Springer, LNCS (in print), 2013.
4. Szymański J., Duch W.: Dynamic Semantic Visual Information Management. Proceedings of the 9th International Conference on Information and Management Sciences, Urumchi, China, 2010, s. 107÷117.

5. Szymański J., Duch W.: Representation of hypertext documents based on terms, links and text compressibility. Neural Information Processing, Theory and Algorithms, Sydney 2010, s. 282÷290.
6. Milgram S.: The small world problem. Psychology today, 2(1), 1967, s. 60÷67.
7. Watts D.: Small worlds: the dynamics of networks between order and randomness. Princeton University Press, 2003.
8. Kleinberg J.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 1999, s. 604÷632.
9. Langville A., Meyer C.: Deeper inside pagerank. Internet Mathematics, 1(3), 2004, s. 335÷380.
10. Deptuła M., Szymański J., Krawczyk H.: Interaktywne wyszukiwanie informacji w dużych kolekcjach danych oparte o zysk informacyjny na podstawie danych z Wikipedii. (w druku), 2012.
11. Szymański J.: Interactive Information Retrieval Algorithm for Wikipedia Articles, Springer, LNCS, 2012, s. 200÷207.
12. Zygmunt A., Koźlak J., Krupczak Ł.: Analiza grup w serwisach społecznościowych. Studia Informatica, Vol. 32, No. 2A (96), Gliwice 2011, s. 365÷376.

Wpłynęło do Redakcji 7 stycznia 2013 r.

Abstract

In the article an approach to identification of relations between categories that organize the repository of documents has been presented. We provide detailed statistic of processed Wikipedia data. We described in detail our approach based on metrics of the category relevance based on similarity measures between articles. The metrics have been used to discover relations between Wikipedia categories. The evaluation of the proposed method indicates that it is possible to reconstruct almost half of already existing associations in category the structure. Also the method allows to introduce new, significant relations between categories. The indentified relations between categories finds many applications especially in information retrieval domain and natural language processing.

Adresy

Julian SZYMAŃSKI: Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Architektury Systemów Komputerowych, ul. Narutowicza 11/12, 80-233 Gdańsk-Wrzeszcz, Polska, julian.szymanski@eti.pg.gda.pl.

Marcin DEPTUŁA: Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Architektury Systemów Komputerowych, ul. Narutowicza 11/12, 80-233 Gdańsk-Wrzeszcz, Polska, marcin.deptula@live.com.

Henryk KRAWCZYK: Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Architektury Systemów Komputerowych, ul. Narutowicza 11/12, 80-233 Gdańsk-Wrzeszcz, Polska, henryk.krawczyk@eti.pg.gda.pl.