INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

Wybrane problemy
Tom 2 (14)

poprzednio

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Pod redakcją
Andrzeja MYŚLIŃSKIEGO

Warszawa 2012

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

Wybrane problemy
Tom 2 (14)

poprzednio

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Pod redakcją
Andrzeja Myślińskiego

Warszawa 2012

Wykaz opiniodawców artykułów zamieszczonych w niniejszym tomie:

Dr hab. inż. Andrzej MYŚLIŃSKI, prof. PAN

Dr hab. inż. Ryszard SMARZEWSKI, prof. KUL

Dr hab. Dominik ŚLĘZAK

Prof. dr hab. inż. Andrzej STRASZAK

Prof. dr hab. inż. Stanisław WALUKIEWICZ

Dr hab. Adam WIERZBICKI

ISBN 9788389475442

KLASYFIKACJA TEMATYCZNA TREŚCI STRON WWW W OPARCIU O STRUKTURĘ WIKIPEDII

Przemysław Kusaj

Studia Doktoranckie IBS PAN
e-mail: *pkusaj@kul.pl*

Streszczenie. Ilość treści dostępnych online jest ogromna. Jej przyrost związany jest z możliwością tworzenia jej przez każdego użytkownika sieci, co zapoczątkowane zostało wraz z rozpoczęciem ery Web 2.0 i stworzeniem narzędzi, które bez specjalistycznej informatycznej wiedzy pozwalają partycypować w dodawaniu treści i wpływać na to, co zawiera. Sprawiło to, że wśród wyników wyszukiwania, oprócz relewantnych stron, wyszukiwarki wyświetlają też strony zawierające poszukiwany termin, ale nie koniecznie z tym terminem tematycznie związane. Problemem staje się szybkie rozpoznanie, czy dana treść jest tą, której poszukujemy, co byłoby uproszczone, gdyby treść była tematycznie sklasyfikowana. Niniejszy artykuł przedstawia propozycję rozwiązania problemu klasyfikacji tematycznej treści znajdujących się na stronach WWW, które niezbędne do klasyfikacji informacje czerpie ze struktury linków wewnętrznych Wikipedii, przez co jest językowo niezależne.

Słowa kluczowe: klasyfikacja tematyczna, data mining, Wikipedia, wektory słów kluczowych

1 WSTĘP

Po 2001 roku, od momentu pęknięcia tzw. „bańki dotcom'ów", kiedy to zwykli użytkownicy przejęli rolę twórców treści, ilość informacji dostępnych w Internecie zaczęła rosnąć w bardzo szybkim tempie. Powstawanie kolejnych, często bardzo bogatych w treść i zróżnicowanych tematycznie witryn, portali i blogów, oraz sklepów internetowych skomplikowały proces odnajdywania informacji w sieci. Naprzeciw temu problemowi wyszły wyszukiwarki internetowe, które za pomocą algorytmów indeksujących i robotów internetowych (web crawler'ów) zbierają informacje o zasobach dostępnych w sieci i starają się je katalogować m.in. na podstawie słów kluczowych znajdujących się w nagłówkach kodu stron. W środowisku o takiej ilości i różnorodności treści rozsianej po ok. 650 milionach aktywnych obecnie stron internetowych pełnią one bardzo istotną funkcję, pomagając wciąż rosnącej rzeszy osób korzystających z Internetu odnaleźć wymagane informacje.

Pojawia się jednak pytanie: jak często otrzymujemy satysfakcjonujące rezultaty po wpisaniu do wyszukiwarki haseł dotyczących interesującego nas tematu? Czy wyszukane artykuły na podanych w wynikach stronach faktycznie będą poświęcone interesującemu nas zagadnieniu, a może zawierają tylko wzmianki na jego temat, a całość tekstu dotyczy czegoś zupełnie innego? Dla przykładu, szansa na to, że zamiast odnośnika do testu konkretnego modelu routera otrzymamy listę linków do sklepów internetowych, gdzie go można nabyć (a gdzie w treści pojawiają się słowa takie jak „opinie" czy „test", lecz brak jest faktycznego testu, a opinie użytkowników nie zostały jeszcze dodane) jest bardzo duża.

Pomimo całej gamy używanych przez wyszukiwarki algorytmów hierarchizujących i oceniających jakość stron WWW, wyniki nie zawsze są zadowalające. W dużej mierze jest to efekt praktyk SEO (Search Engine Optimization), czyli optymalizacji wyników wyszukiwania stron WWW, prowadzonych przez twórców stron oraz przez komercyjne firmy, których celem jest zmanipulowanie wyników pracy algorytmów hierarchizujących tak, by link do danej strony był jak najwyżej w wynikach wyszukiwania dla danej grupy haseł.

Poniższy artykuł przedstawi koncepcję klasyfikacji tematycznej dowolnych treści znajdujących się na stronach WWW, która może okazać się pomocna w rozwiązaniu problemu trafnej i szybkiej oceny stron pod względem zgodności z interesującym nas zagadnieniem. Koncepcja ta będzie testowana, jako część przygotowywanej pracy doktorskiej.

2 ZARYS PROPONOWANEGO ROZWIĄZANIA

Po zapoznaniu się z technikami używanymi przy optymalizacji stron można wywnioskować, że rozpoznanie strony jako traktującej o danej tematyce w niewielkim stopniu zależy od tego, jakie treści zawiera. Płynie z tego wniosek, że algorytmy poddające ocenie tematykę strony niezbyt dokładnie (jeśli w ogóle) sprawdzają jej faktyczną zawartość.

Proponowane rozwiązanie bazuje właśnie na sprawdzeniu, czego dotyczy główna treść każdej pojedynczej strony, a właściwie artykułu na danej podstronie, i nie bierze po uwagę żadnych dodatkowych informacji wplatanych przez developera w jej kod jako meta dane czy nagłówki. Mowa tu o podstronach, ponieważ przyjmujemy, że portal czy witryna WWW może być zróżnicowana tematycznie, zatem jednoznaczna ocena ukierunkowania tematycznego danej strony jako całości jest procesem problematycznym i złożonym. Dodatkowo, opieramy się na obserwacji wyników działania wyszukiwarek, gdzie również podawane są listy podstron, a nie

adresy do stron startowych poszczególnych witryn. Przyjmijmy zatem, że w dalszej części artykułu, gdy mowa o treści na stronie WWW, chodzi tak naprawdę o podstronę (pojedynczy dokument HTML) i znajdującą się na niej treść (np. artykuł).

Analiza głównego tekstu na stronie pozwala na utworzenie wektora słów kluczowych (w dalszej części niniejszego artykułu określany skrótowo jako WSK) dla tego tekstu. Wektor taki zawierałby terminy charakterystyczne dla analizowanego tekstu wraz częstością ich wystąpień. Daje to możliwość podjęcia próby jego klasyfikacji tematycznej, poprzez wyznaczenie miar podobieństwa tego wektora do wzorcowych WSK charakteryzujących poszczególne zakresy (kategorie) tematyczne.

Wyznaczenie wzorcowych WSK, które mogłyby jednoznacznie charakteryzować każdą kategorię tematyczną, wymaga rzetelnych źródeł, które merytorycznie wyczerpują (w miarę możliwości) poszczególne tematy. Pojedynczym źródłem, w którym odnaleźć można informacje na prawie każdy temat jest Wikipedia - otwarta encyklopedia online (choć należy zauważyć, że ilość opisanych tematów, a także poziom wyczerpania danego tematu oraz ilość informacji o tematach pokrewnych zależy od ilości artykułów w ogóle, co z kolei związane jest ze stopniem rozwoju danej wersji językowej Wikipedii). Do zalet wykorzystania Wikipedii w procesie konstruowania wzorcowych WSK należy zaliczyć to, że każda jej wersja językowa jest dynamicznie rozwijana i stale poszerzana o kolejne hasła, treści są regularnie aktualizowane i weryfikowane, a cała zawartość poszczególnych jej wersji językowych udostępniana jest w postaci plików XML (tzw. *dump*) na zasadach licencji otwartych *Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA)* i *GNU Free Documentation License (GFDL)* [1]. Niedogodnością natomiast jest jej chaotyczna struktura. Podobnie jak treści artykułów, tak i ich przypisania do kategorii tematycznych Wikipedii zależne są od użytkowników. Zarówno jedne jak i drugie są cyklicznie weryfikowane, jednak Wikipedia dopuszcza przynależność hasła do wielu kategorii i podkategorii jednocześnie. Dokumentacja Wikipedii nakazuje też, aby w pisanym artykule użytkownik utworzył hiperłącze dla każdego terminu i frazy, które związane są z jakimkolwiek znaczącym zagadnieniem, do artykułu, który traktuje o tym zagadnieniu. Nawet jeśli taki artykuł jeszcze nie istnieje, łącze jest tworzone, prowadzi natomiast do strony zachęcającej do napisania odpowiedniego artykułu. To sprawia, że graf relacji pomiędzy jej artykułami jest bardzo skomplikowany i w takiej formie nie może ona posłużyć do skonstruowania wzorcowych WSK, które byłyby w stanie jednoznacznie charakteryzować kategorie tematyczne.

Problem ten da się jednak rozwiązać bazując na sposobie analizy struktury Wikipedii opisanym przez Lizorkina et al. w [2], opartym na algorytmie *Girvana-Newmana (community detection algorithm)* do wyznaczania mocno powiązanych ze sobą grup węzłów (podsieci) w rozbudowanych i skomplikowanych sieciach (*complex networks*) [4]. W kontekście Wikipedii będziemy zatem mówić o wyznaczaniu klastrów tematycznych, czyli wyodrębnianiu grup artykułów ściśle powiązanych ze sobą wspólną tematyką przy założeniu, że każdy artykuł może należeć tylko i wyłącznie do jednej grupy, i to tej, do której tematycznie jest mu najbliżej.

3 PRZYGOTOWANIE KORPUSU DO WYZNACZANIA KATEGORII TEMATYCZNYCH

Wikipedia jest encyklopedią otwartą i w związku z tym każdy może partycypować w jej tworzeniu i rozwoju poprzez pisanie artykułów, sprawdzanie ich, zgłaszanie nieścisłości bądź błędów, itp. Jak już nadmieniono powyżej, to twórca artykułu Wikipedii przypisuje dany artykuł do konkretnej kategorii oraz, poprzez umieszczanie konkretnych hiperłącz w treści, wskazuje relacje pomiędzy tym artykułem a innymi artykułami, tematami, kategoriami. Zatem można uznać, że cała struktura Wikipedii, podział artykułów na kategorie i zależności semantyczne pomiędzy artykułami określane są w całości przez użytkowników. Biorąc pod uwagę dynamiczny rozwój tej encyklopedii i wciąż rosnącą ilość artykułów (w polskiej Wikipedii obecnie ok. 887 000, w anglojęzycznej natomiast ponad 3 970 000) prawidłowe określenie wszystkich połączeń tematycznych artykułu z innymi dla osoby piszącej ten artykuł może być problematyczne. Jak słusznie zauważyli Lizorkin et al. [2], fakt ten sprawia, że zależności semantyczne w obrębie Wikipedii są niepełne i nie zawsze trafne. Znalezione w anglojęzycznej Wikipedii, podane przez nich przykłady pokazują, na jakiego typu problemy można natknąć się analizując graf połączeń artykułów w Wikipedii. Np. *Domestic pig* $\subset$ (należy do kategorii) *Pork*, co nie wydaje się być do końca poprawnym przypisaniem do kategorii. Kolejnym problemem są często występujące w połączeniach cykle takie, jak następujący: *The Beatles* $\subset$ *Apple Records* $\subset$ *Apple Corps* $\subset$ *The Beatles*.

Wikipedia w takiej formie, czyli z tego typu nieścisłościami w obrębie zależności pomiędzy artykułami, nie może być użyta do konstrukcji wektorów słów kluczowych, które miałyby jednoznacznie charakteryzować poszczególne kategorie tematyczne. Kategorie tematyczne oraz przypisane do nich grupy artykułów muszą zostać jasno określone. Aby uprościć graf zależności wewnętrznych Wikipedii, autorzy [2], patrząc na nie-

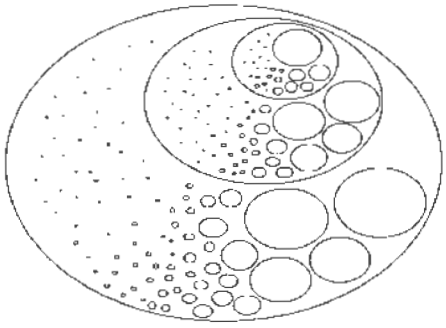
go przez pryzmat skomplikowanej sieci, zaproponowali, aby wyznaczyć w jego obrębie tzw. communities, co można rozumieć jako podgrafy, podsieci lub klastry (co w przypadku podziału tematycznego zdaje się być najtrafniejszym określeniem). Graf zależności Wikipedii, po wyznaczeniu klastrów, w strukturze przypomina sieć komputerową z jej podsieciami. Każdy klaster tematyczny składa się z pewnej liczby ściśle powiązanych ze sobą tematycznie artykułów, a relacje pomiędzy poszczególnymi tematami reprezentowane są przez pojedyncze połączenia pomiędzy konkretnymi klastrami. Użytym algorytmem wykrywania klastrów był algorytm Girvan-Newman, który analizuje połączenia pomiędzy węzłami grafu i dzieli graf na mniejsze podgrafy (moduły) w oparciu o miarę zwaną modułowością (*modularity*). Miara ta porównuje moc połączeń pomiędzy węzłami w grafie podzielonym na moduły (czyli klastry) do tej w grafie o losowych połączeniach pomiędzy węzłami. O grafie wykazującym wyraźny podział na klastry możemy mówić, gdy wartość modułowości wynosi pomiędzy 0,3 a 0,7. W przypadku Wikipedii połączeniami branymi pod uwagę są linki wewnętrzne umieszczane w artykułach, czyli tylko te prowadzące do innych stron Wikipedii, i to tylko te znaczące, tzn. linki do stron kategorii, do strony głównej, te pomiędzy artykułami z tej samej kategorii, co rozpatrywany artykuł, linki z sekcji *„Zobacz także:”* oraz linki wzajemne (artykuł A posiada link do artykułu B, a artykuł B posiada link do A).

Analiza anglojęzycznej Wikipedii przeprowadzona przez Lizorkina et al. w 2008 r. wykazała, że na bazie jej grafu, na który składało się wówczas 4,1 mln węzłów (wszystkich stron, nie tylko artykułów), po zastosowaniu algorytmu wykrywania klastrów, dało się ustalić 2 038 klastrów o bardzo wysokim stopniu modułowości wynoszącym 0,63 - oznacza to, że graf Wikipedii da się podzielić na klastry tematyczne, które będą ze sobą odpowiednio połączone. Spójność tematyczna wewnątrz klastrów jednak, jak wykazano, jest tym mniejsza im więcej artykułów składa się na klaster.

Posługując się metodą *Wikipedia Link-based Measure* (WLM) do obliczania miar pokrewieństwa semantycznego pomiędzy terminami na bazie zawartych w treści artykułów słów „kotwiczących” (*anchor-texts*) i powiązanych z nimi linkami prowadzącymi do innych artykułów wewnątrz Wikipedii, opracowaną przez Milne et al. [3], przeprowadzili oni ewaluację pokrewieństwa semantycznego (*semantic relatedness*) artykułów Wikipedii i wykazali, że pokrewieństwo to pomiędzy losowo wybranymi artykułami bliskie jest 0, natomiast w małych klastrach do 50 artykułów wynosi średnio 35%, a w tych do 1000 artykułów średnio 25%. Jest to dowodem

na to, że artykuły w małych klastrach są ze sobą dość mocno powiązane i wszystkie dotyczą wspólnego tematu. Należy również wspomnieć, że metoda WLM, dzięki temu, że pracuje na linkach, reprezentujących je w treści „anchor-text'ach" i terminach, do których prowadzą, rozwiązuje problemy związane z polisemią, czyli istnieniem wielu terminów opisujących ten samo zjawisko czy przedmiot, oraz z przypadkami, kiedy dany termin ma różne znaczenia w zależności od kontekstu, i może być użyta w każdej dowolnej wersji językowej Wikipedii. Ponadto, metoda ta pozwala ustalić pokrewieństwo semantyczne terminów niższym kosztem niż poprzez analizę tekstową.

Rozkład wielkości klastrów po jednym przebiegu algorytmu po oryginalnym grafie Wikipedii przypomina rozkład *power law*, z kilkoma wielkimi klastrami (powyżej 300 tys. artykułów) i wieloma (99%) małymi (poniżej 2 tys. węzłów) [2]. Jeżeli zatem po jednym przebiegu algorytmu klastry są zbyt duże, by móc mówić o wysokim stopniu pokrewieństwa semantycznego, należy przeprowadzić wykrywanie klastrów drugiego poziomu w tych klastrach (podgrafach) wyznaczonych w pierwszym przebiegu algorytmu, które nie wykazują dostatecznego stopnia wewnętrznego pokrewieństwa i kontynuować ten proces rekurencyjnie na kolejnych poziomach, aż do uzyskania hierarchicznej struktury o odpowiednio małych klastrach na najniższych poziomach tej struktury, które charakteryzowałyby się satysfakcjonującym wewnętrznym stopniem pokrewieństwa semantycznego i spójnością tematyczną pomiędzy ich artykułami. Lizorkin et al. [2] sugerują, aby wyznaczać pod-klastry kolejnych poziomów kierując się wartościami modułowości, aż do uzyskania odpowiednio semantycznie połączonych ze sobą, wewnętrznie spójnych grup artykułów.



Rys. 1. Wizualizacja hierarchicznej struktury klastrowej. Źródło:[2]

Aby określić wspólny temat artykułów każdego z klastrów uzyskanych przez powyżej opisane podziały w [2] posłużono się algorytmem *PageRank*, który aplikowano na każdy klaster. Wykazano, że, jeżeli klaster tematyczny został odpowiednio wyznaczony, wśród 3 najwyższych wyników wskazanych przez algorytm znajduje się strona kategorii Wikipedii (strona z działu *„Kategorie”*), w niższych wynikach natomiast nie ma innych stron tego typu. Jeżeli w wynikach działania *PageRank*'a znajdzie się więcej stron kategorii, należy zastanowić się nad dalszym podziałem klastra na pod-klastry. Jeżeli w najwyższych wynikach nie znajdzie się żadna strona kategorii, można przyjąć, że artykuł znajdujący się na najwyższej pozycji (a właściwie jego tytuł), jest artykułem centralnym i określa tematykę klastra.

Artykuł centralny można też ustalić w oparciu o algorytm *HITS* opracowany przez J. Kleinberga [6], służący do wyznaczania autorytatywnych stron WWW z danego zbioru (np. ze wszystkich wyszukanych stron dla konkretnego zapytania w wyszukiwarce internetowej), a tym samym tworzenia rankingu stron wg ich zgodności z poszukiwanym wyrażeniem. Wyznaczone klastry zdają się spełniać wszystkie 3 warunki, które wg autora algorytmu musi spełniać każdy zbiór bazowy stron WWW, dla którego ma być tworzony ranking:

- każdy klaster skupia artykuły semantycznie ze sobą pokrewne, a więc zestaw jest bogaty w dokumenty związane z poszukiwanym zagadnieniem (tematem głównym);
- klaster składa się z określonej liczby mocno powiązanych ze sobą poprzez wzajemne hiperłącza artykułów, więc najprawdopodobniej zawiera też stosunkowo dużą ilość autorytatywnych artykułów (do których linkuje duża ilość artykułów w obrębie klastra);
- sam zbiór jest stosunkowo niewielki (w porównaniu z całością Wikipedii, gdyż zawiera tylko artykuły spójne ze sobą tematycznie).

Pomimo, że algorytm w niekontrolowanych warunkach sieci WWW wykazuje podatność na manipulację, np. poprzez tworzenie tzw. linków nepotycznych, które zawyżają autorytet danej strony, w tym przypadku nie ma to znaczenia, ponieważ każdy element zbioru jest tworzony zgodnie ze standardami określonymi w dokumentacji Wikipedii i weryfikowany przez społeczność współtwórców w taki sposób, by te standardy spełniać.

Dzięki powyższemu rozwiązaniu, niezależnie od wybranego algorytmu ustalania artykułu centralnego, otrzymuje się korpus złożony z wewnętrznie spójnych tematycznie klastrów z jasno określoną kategorią tematyczną.

Na bazie każdego z klastrów można zatem wyznaczyć wzorcowy wektor słów kluczowych charakterystyczny dla danego tematu.

Warto też zauważyć, że powyższe rozwiązanie jest uniwersalne i może być aplikowane na grafie każdej wersji językowej Wikipedii, ponieważ przy podziałach rozpatruje tylko połączenia pomiędzy artykułami (linki wewnętrzne zawarte w treści artykułów) a nie same treści, przez co jest niezależne od języka, w którym pisane są artykuły.

4 WYZNACZANIE WEKTORÓW SŁÓW KLUCZOWYCH

4.1 Wyznaczanie wzorcowych wektorów słów kluczowych identyfikujących kategorie tematyczne

Bazując na regule, że wszystkie występujące w treści artykułu terminy czy frazy nawiązujące do zagadnień znaczących dla tematu tego artykułu muszą jednocześnie tworzyć hiperłącza do odpowiednich artykułów opisujących te zagadnienia, można wysnuć wniosek, że każde znaczące (kluczowe) dla danego artykułu słowo, jest oznaczone w treści jako odnośnik hipertekstowy, co sprawia, że identyfikacja słów kluczowych artykułu jest uproszczona.

Zgodnie z założeniem opisywanej tutaj koncepcji, nie chodzi jednak tylko o słowa kluczowe poszczególnych artykułów, a o te właściwe dla danego tematu. Jeżeli zatem przyjmujemy, że każdy wyznaczony klaster zawiera wszystkie artykuły dotyczące jednego, konkretnego tematu, wówczas traktując ten klaster jako jeden duży i tematycznie spójny artykuł, można wyznaczyć wzorcowy WSK dla reprezentowanej przez ten klaster kategorii tematycznej na bazie słów kluczowych wszystkich artykułów wchodzących w skład klastra. Można też uznać, że po wyznaczeniu klastrów, terminy będące słowami kluczowymi w danym klastrze to wyrażenia o jednoznacznie określonym znaczeniu właściwym dla kontekstu tematycznego tego klastra. Te same terminy mogą oczywiście znajdować się w innych WSK innych klastrów tematycznych, ale w większości przypadków będą one występowały w odmiennych znaczeniach. Jeśli jednak będą występować w tym samym znaczeniu, co w innym WSK, to najprawdopodobniej w innych ilościach, co z kolei będzie mieć wpływ na wagę tych terminów w określaniu tematyki danego klastra.

Zaletą oparcia tego rozwiązania na analizie hiperłącz jest również to, że słowa składające się na wielowyrazowe wyrażenia kluczowe nie ulegają rozdzieleniu i nawet jeśli w treści występują w formie odmienionej przez przypadki to, jako hiperłącza prowadzą do odpowiadających sobie

artykułów, których tytuły zapisane są w formie mianownikowej. Ponadto, jeżeli istnieją oficjalne skróty lub alternatywne nazwy tych wyrażeń (czy nawet jednowyrazowych terminów), przypadki takie powinny automatycznie ulec rozwiązaniu, ponieważ do każdego terminu, który takowe posiada prowadzą łącza od tzw. stron ujednolicających znajdujących się na Wikipedii.

Przyglądając się niektórym artykułom Wikipedii można zauważyć, że w części z nich ilość słów kluczowych i wyrażeń może być niewielka. Mimo to, jeżeli rozpatrujemy wszystkie artykuły danej grupy tematycznej jako całość, zbiór słów i wyrażeń kluczowych ze wszystkich artykułów wchodzących w skład klastra powinien dać wektor reprezentatywny dla określonej tematyki.

Mając listę słów kluczowych należy następnie utworzyć wektor TF-IDF bazując na modelu *Bag of Words* (BOW, zwany też inaczej *Vector Space Model*) [5]. Byłby to wektor o ilości wymiarów takiej, jak liczba wszystkich słów kluczowych korpusu, zawierający ilości wystąpień poszczególnych słów i wyrażeń kluczowych w danym klastrze ważone ilością wszystkich słów w tym klastrze i normalizowany odwrotną częstotością występowania klastrów zawierających poszczególne słowa kluczowe składające się na wektor w stosunku do wszystkich klastrów tematycznych wyznaczonych w obrębie Wikipedii, tworzony wg następujących wzorów:

- Wylistowanie wszystkich relewantnych słów kluczowych ze wszystkich artykułów klastra i oszacowanie częstotliwości ich występowania w klastrze $c_n \subset C$, gdzie C to zbiór wszystkich klastrów Wikipedii

$$TF_{C_n} = \frac{KW_i}{T}$$

  KW_i -ilość wystąpień słowa kluczowego *i* ; $i \subset \{0, 1, \ldots., T-1\}$
  T - ilość wszystkich słów we wszystkich artykułach $\subset C_n$

- Normalizacja przez odwrotność częstości występowania klastrów zawierających dane słowo kluczowe

$$IDF_{C_n} = \frac{|W|}{|C|}$$

  $|W|$ - ilość wszystkich klastrów Wikipedii
  $|C|$ - ilość klastrów zawierających słowo kluczowe i

Wszystkie słowa i wyrażenia kluczowe wykryte we wszystkich klastrach stworzą korpus dla klasyfikacji tematycznej treści. Następnie ze wszystkich wzorcowych WSK powinna zostać utworzona macierz, która gromadziłaby ważone częstości występowania w poszczególnych klastrach wszystkich słów i wyrażeń kluczowych korpusu.

4.2 Wyznaczanie wektora słów kluczowych treści na stronie WWW

Określanie wektora słów kluczowych treści dowolnej strony WWW wymaga upewnienia się, że rozpatrywany jest tylko i wyłącznie tekst znaczący dla tematyki strony. Niezbędnym jest odseparowanie wszelkich innych treści, które mogłyby wpłynąć na proces klasyfikacji tematycznej, a tym samym i trafność tej klasyfikacji, tj. treści reklamowych, elementów stałych strony (np. wspólnych dla wszystkich stron danego portalu), linków, opisów zdjęć (ponieważ te mogą nie mieć związku z artykułem na stronie), itp. Pomocna w celu określenia głównej treści strony WWW będzie analiza źródła (kodu) strony.

Gdy tekst główny zostanie odseparowany konieczne jest jego odpowiednie przygotowanie do analizy pod względem poszukiwania słów i wyrażeń kluczowych. Niezbędne będzie przeprowadzenie następujących operacji:

- odseparowanie tzw. „stopwords", czyli najczęściej występujących w języku słów nie niosących z sobą żadnych istotnych treści. Lista takich słów dla języka polskiego znajduje się na jednej ze stron Wikipedii pod adresem [7].
- wyselekcjonowanie i zachowanie wyrażeń kluczowych składających się z więcej niż jednego wyrazu (co może okazać się kłopotliwe, pod warunkiem, że nie występują one jako odnośniki hipertekstowe).
- *„Stemming"* słów, tzn. odcięcie ich o końcówek, tak by uzyskać temat każdego z wyrazów, czyli jego fragment nie podlegający odmianie. Aby jednak nie utracić informacji należy uważnie przeprowadzać ten proces, jeśli rozpatruje się wielowyrazowe wyrażenia kluczowe [5]. Zdecydowanie jest to problem, którego rozwiązanie należy dogłębnie przemyśleć. Znalezione rozwiązania dotyczą tego typu wyrażeń w języku angielskim - z uwagi na fakt, że nie odmieniają się one w żaden sposób, zaleca się, aby nie przeprowadzać stemming'u. Przykładem niech będzie wyrażenie „machine learning" - stemming sprawiłby, że słowo „learning" zostałoby skrócone do „learn", a to może zaburzyć statystykę występowania tego wyrazu w treści, a tym samym i miarę

podobieństwa wektorów słów kluczowych. W języku polskim może to jednak wyglądać zupełnie inaczej, dlatego też konieczne jest przemyślenie i określenie sposobów działania z wyrażeniami wielowyrazowymi.
- Odrzucenie słów, które występują na tyle rzadko w rozpatrywanym tekście, że zdecydowanie nie będą miały wpływu na klasyfikację treści.

Mając tak przygotowany tekst można rozpocząć wyszukiwanie składowych dla WSK.

Pierwszym etapem powinno być wyszukanie i zliczenie wielowyrazowych wyrażeń, które występują w tekście. Do identyfikacji tego, co może być takowym wyrażeniem posłuży korpus słów kluczowych przygotowany na bazie Wikipedii, ponieważ zawiera on wszystkie te, które mogą przysłużyć się do klasyfikacji tematycznej treści. Konieczne będzie opracowanie algorytmu, który bazując na korpusie zidentyfikuje w treści strony WWW te wyrażenia. Wzorem może być przytaczany w [8] moduł o nazwie *Terminator*, który pracuje w oparciu o struktury gramatyczne języka angielskiego, które na bieżąco wyszukuje dla napotkanych w treści słów w słowniku języka angielskiego dostępnym online.

Po odnalezieniu wszystkich relewantnych wyrażeń kluczowych należałoby je odseparować od pozostałej treści artykułu, by wyrazy wchodzące w ich skład nie zawyżały częstości występowania wyrazów w trakcie zliczania pojedynczych słów kluczowych. Po tej operacji należy zidentyfikować i obliczyć częstość występowania pojedynczych słów kluczowych, które również znajdują się w korpusie.

Na tym etapie można przystąpić do konstrukcji WSK. Podobnie, jak w przypadku wzorcowych WSK, wektor ten składać się będzie z wartości częstości występowania danego słowa kluczowego w treści strony WWW dzielonej przez ilość wszystkich słów w rozpatrywanym tekście i normalizowanej przez odwrotność ilości klastrów zawierających to słowo w stosunku do ilości wszystkich klastrów w Wikipedii (ponieważ ciągle traktujemy każdy z klastrów jako jeden duży artykuł). Jest to w pewnym sensie próba potraktowania treści na stronie jako osobnego klastra Wikipedii.

5 METODA KLASYFIKACJI TEMATYCZNEJ W OPARCIU O WEKTORY SŁÓW KLUCZOWYCH

Wyznaczony WSK strony WWW należy poddać analizie jego bliskości w stosunku do poszczególnych wzorcowych WSK przygotowanych na podstawie klastrów tematycznych Wikipedii.

Aby uznać, że wektor słów kluczowych danej treści strony WWW wskazuje na konkretną tematykę, należy obliczyć podobieństwo cosinusowe (*cosine similarity*) tego wektora w stosunku do wzorcowych WSK charakteryzujących różne kategorie tematyczne (tj. cosinus kąta pomiędzy rozpatrywanym WSK i każdym z wzorcowych WSK, co da miary podobieństwa tych wektorów). Jeżeli miara ta wynosi 1 to daje to jednoznaczną odpowiedź, że wektory są identyczne, 0 dowodzi wzajemnej niezależności wektorów, -1 natomiast wykazuje, że wektory są skrajnie różne. Wartość cosinusowa najbliższa jedności da odpowiedź, do którego wzorcowego WSK najbardziej podobny jest WSK strony WWW, a to z kolei pozwoli określić do jakiej kategorii tematycznej najbliżej jest analizowanej treści.

6 ZALETY I WADY ROZWIĄZANIA

Wikipedia to encyklopedia dynamicznie rozwijana - ciągle dodawane są nowe artykuły, a informacje na istniejących są cyklicznie weryfikowane i aktualizowane. Oznacza to, że wiedza zawarta w Wikipedii poszerzana i uściślana jest z każdą zachodzącą zmianą. Zaletą wynikającą z powyższego faktu jest to, że w miarę poszerzania się encyklopedii o kolejne informacje ilość i jakość słów kluczowych charakteryzujących zagadnienia i tematy będzie rosnąć. Cztery razy do roku tworzony i publikowany jest tzw. dump wszystkich artykułów w postaci pliku XML. Oznacza to, że, jeżeli rozwiązanie jest poprawne, w każdym kwartale, korzystając z najaktualniejszej wersji dump'u możliwe ponowne wyznaczenie klastrów i wzorcowych WSK, które jeszcze dokładniej charakteryzować będą kategorie tematyczne. Zatem, z kwartału na kwartał wyniki klasyfikacji tematycznej treści stron WWW w oparciu o korpus powinny być coraz trafniejsze.

Wikipedia prowadzona jest w wielu wersjach językowych, co w przyszłości daje możliwość zastosowania proponowanego rozwiązania do stron pisanych w każdym z tych języków, ponieważ samo podejście jest uniwersalne. Bazuje ono na analizie hiperłącz istniejących pomiędzy artykułami, a nie samej treści, co sprawia, że jest językowo niezależne.

Podobnie, algorytm wyznaczający WSK analizowanych w celu klasyfikacji treści stron działać będzie w oparciu o korpus zbudowany na bazie konkretnej wersji językowej Wikipedii i, czerpiąc z niego informacje, w sposób zautomatyzowany przeszukiwać będzie treść w poszukiwaniu zawartych w korpusie słów i wyrażeń kluczowych. Obsługiwany przez niego język w pełni zależeć będzie od zestawu słów kluczowych dostarczonych przez korpus.

W przypadku niektórych języków przygotowanie treści artykułów do ekstrakcji słów kluczowych, może okazać się problemem - chodzi głównie o kwestie związane ze stemmingiem. Rozwiązanie ich będzie wymagało dostarczenia algorytmom informacji na temat specyfiki i gramatyki języka (np. kwestie budowy i deklinacji wyrazów).

Należy też nadmienić, że proponowane rozwiązanie ukierunkowane jest na klasyfikację treści w formie artykułów dostępnych online, przy założeniu, że artykuł w całości znajduje się na danej, pojedynczej stronie. Klasyfikacja przeprowadzona tym sposobem z pewnością nie da trafnej odpowiedzi co do tematyki treści, jeżeli na stronie znajdować się będzie więcej niż jeden spójny i jednolity tekst, jak to ma miejsce w przypadku blogów, gdzie na jednej stronie można znaleźć kilka niezależnych od siebie tematycznie tekstów. Podobnie treści o charakterze prywatnym, mocno nacechowane emocjonalnie mogą dać nietrafne wyniki klasyfikacji.

7 PERSPEKTYWY ROZWOJU ROZWIĄZANIA

W najbliższym czasie planowana jest realizacja opisanego powyżej rozwiązania problemu klasyfikacji treści na stronach WWW. Testy wykażą, na ile jest to poprawny sposób i jak trafnie klasyfikacja jest przeprowadzana.

Jeżeli zaproponowane rozwiązanie da spodziewane efekty, to kolejnym etapem rozwoju będzie próba weryfikacji sklasyfikowanych treści pod względem ich wiarygodności.

Literatura

1. http://en.wikipedia.org/wiki/Wikipedia:Database_download
2. Lizorkin D., Medelyan O., Grineva M. (2009) Analysis of Community Structure in Wikipedia (Poster) in *Proceedings of WWW 2009*, Madrid.
3. Milne D, Witten I (2008) An effective, low-cost measure of semantic relatedness obtained from Wikipedia links in *Wikipedia and AI workshop at the AAAI*.
4. Newman M.E.J., Clauset A., Moore C. (2004) Finding community structure in very large networks, *Physical Review E*, 70:066111.
5. Wang P. , Domeniconi C. (2008) Building Semantic Kernels for Text Classification using Wikipedia in *14th ACM SIGKDD*, New York, 713-721.
6. Kleinberg J. (1998) Authoritative sources in a hyperlinked environment in *Proc. 9th Ann. ACM-SIAM Symp. Discrete Algorithms*, ACM Press, New York, 668-677.
7. http://pl.wikipedia.org/wiki/Wikipedia:Stopwords
8. Kazi Z., Ravin Y. (2000) Who's Who? Identifying Concepts and Entities across Multiple Documents in *Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE*, 1-7.

THEMATICAL CLASSIFICATION OF WEBPAGE CONTENT BASED ON WIKIPEDIA STRUCTURE

Abstract. The amount of content available on-line is enormous. Its growth is related to the fact that nowadays every user can create content, which was made possible with the introduction of various content creation tools, that do not require any specific computer knowledge, at the beginning of Web 2.0 era. As a result searching through the content for a certain term in certain context became more difficult. When using an on-line search engine, one gets a list of various web pages containing the required term but only a few of them are appear to be relevant regarding the context. The rest of the pages simply contain the term but not always are thematically related to it. The problem that the user faces is to quickly identify if a given text is the relevant one. This could be solved if the content available on-line was classified thematically. This article presents a possible solution to the problem of thematic classification of website content that derives all the information necessary for this purpose from the internal-link structure of Wikipedia, what makes it language-independent.

ISBN 9788389475442