ISSN 1232-3829
tts
10/2013
Technika Transportu Szynowego koleje ▪ tramwaje ▪ metro

GŁADYSZ Anna

WYBRANE METODY AUTOMATYCZNEJ IDENTYFIKACJI SŁÓW KLUCZOWYCH

Streszczenie

Rozwój społeczeństwa informacyjnego oraz technologii informatycznych pociągnął za sobą w sposób naturalny powstanie zautomatyzowanych systemów wspomagających wyszukiwanie i porządkowanie informacji. W nadmiarze informacji przechowywanych w dokumentach tekstowych dużego znaczenia nabiera możliwość automatycznego identyfikowania słów kluczowych. Artykuł rozpoczyna cykl poświęcony badaniu metod algebraicznych wykorzystywanych do identyfikacji słów kluczowych w polskojęzycznych tekstach naukowych.

WSTĘP

Fundamentalnymi czynnikami konstytuującymi obecną rzeczywistość jest zalew informacji, rosnące tempo życia i rosnąca liczba zmian [2, s. 10-12]. Gospodarka, nauka, szeroko pojęte zarządzanie, logistyka oraz każda inna dziedzina tworzą olbrzymie ilości danych, gromadzonych następnie w pamięciach komputerów. Możliwości techniczne są olbrzymie, w komputerach gromadzone są gigantyczne ilości danych, które bez większego trudu mogą zostać dostarczone w dowolne miejsce świata. Dane zbierane i gromadzone są niemal wszędzie, często zupełnie automatycznie, bez udziału człowieka.

Nadmiar dostępnych danych okazuje się trudny do ogarnięcia i analizy, co stanowi poważny o ile nie najpoważniejszy problem pojawiający się na styku człowiek — komputer. Człowiek dysponuje umysłem daleko doskonalszym od najlepszej maszyny, jednak jego możliwości percepcji są mocno ograniczone. Komputer, pomimo dużej objętości pamięci i szybkości działania potrafi w znikomym stopniu zrozumieć człowieka i świat (odgadnąć jego zamiary, intencje, czy potrzeby). Nie dziwi zatem, że to właśnie rozumienie języka naturalnego przez maszyny wydaje się być najbardziej ambitnym, a zarazem najodleglejszym celem jaki może osiągnąć informatyka. Już Alan Turing proponując swój słynny test, badający czy dana maszyna może być uznana za inteligentną, jako kryterium przyjął rozumienie języka naturalnego [15]. Pomijając kwestię sprawdzenia inteligencji maszyny, to stworzenie komputera, czy też algorytmu rozumiejącego język naturalny stanowiłoby wielki przełom w badaniach, gdyż posiadałby on umiejętność analizy repozytorium wiedzy budowanego przez ludzkość przez ostatnie kilka tysięcy lat. Cała bowiem informacja, wiedza i doświadczenie ludzkości zapisywana i przekazywana jest właśnie w języku naturalnym.

Z uwagi na to, że ostatecznym odbiorcą danych i informacji jest człowiek, dla niego dokonuje się wszelkiego rodzaju przetwarzania danych. Problemem, bowiem nie jest efektywne gromadzenie i przetwarzanie danych, lecz zdolność interpretacji i wyciągania użytecznych wniosków [3, s. 37-54].

Z praktycznego punktu widzenia szczególnego znaczenia nabiera automatyczne identyfikowanie słów kluczowych obecnych w treści dokumentu. Istotność problemu

wymagającego szerokiej analizy i badań empirycznych stanowiła przesłanki do rozpoczęcia cyklu artykułów związanych z omawianą tematyką. Celem artykułu jest teoretyczna analiza i empiryczna weryfikacja przydatności użycia wybranych metod identyfikacji słów kluczowych w naukowych tekstach polskojęzycznych.

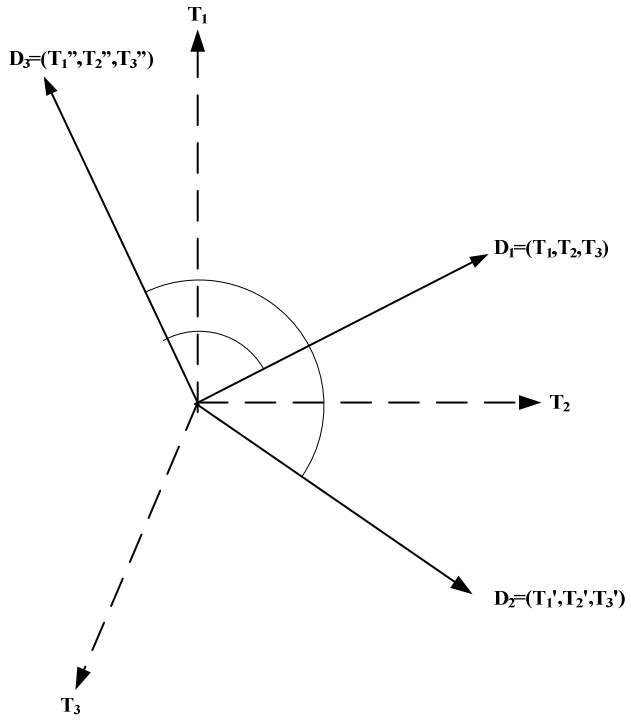
1. REPREZENTACJA INFORMACJI ZAWARTEJ W DOKUMENTACH TEKSTOWYCH

Podstawową metodą reprezentacji informacji pochodzących z dokumentów tekstowych stosowaną w eksploracyjnej analizie tekstu jest model przestrzeni wektorowej.

1.1. Model przestrzeni wektorowej

Model przestrzeni wektorowej (ang. Vector Space Model) został początkowo zaproponowany przez Gerarda Saltona w dziedzinie wydobywania informacji [14]. Jest on algebraicznym modelem reprezentacji dokumentów tekstowych jako wektorów [13, 7, 5, 6].

Rozważana jest przestrzeń dokumentów składająca się z D_i dokumentów, każdy zidentyfikowany przez jeden lub więcej indeksów termów[1] t_j, termy mogą mieć przypisane wagi w zależności od ich znaczenia bądź ograniczone wagami 0 i 1. Typowa trójwymiarowa przestrzeń indeksowa ukazana została na poniższym rysunku (Rys. 1.), w którym każdy dokument jest identyfikowany przez trzy różne termy.



Rys. 1. Reprezentacja wektorowa przestrzeni dokumentu.
Źródło: opracowanie własne na podstawie [14]

Trójwymiarowy przestrzeń może być rozszerzona na *t* wymiarów, gdy występuje *t* różnych indeksowanych termów. D_i reprezentuje *t*-wymiarowy wektor:

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{il}) \tag{1}$$

gdzie d_{ij} stanowi wagę *j*-tego termu.

[1] Pod pojęciem termu rozumie się pojedyncze słowo, lub tez grupę słów, których znaczenie jest całkiem inne niż pojedynczych słów składających się na grupę np. izba, „Wysoka Izba"

Każdy wymiar odpowiada jednemu termowi. Jeżeli term występuje w dokumencie, jego wartość w wektorze jest niezerowa. Na przykład, jeśli wektor *d* reprezentuje dokument *j* w t-wymiarowej przestrzeni, komponent *k* wektora *d*, gdzie $k \in 1 \ldots t$, przedstawia stopień związku pomiędzy dokumentem *j* a termem odpowiadającym wymiarowi *k* w przestrzeni. Ten stopień związku jest lepiej wyrażany jako macierz częstości (ang. *term frequency matrix*) o rozmiarach $t \times d$, gdzie *t* jest numerem unikalnego termu, a *d* jest numerem dokumentu [17]. Element a_{ij} macierzy częstości jest liczbowym przedstawieniem związku pomiędzy termem *i* a dokumentem *j*, reprezentuje liczbę wystąpień termu *i* w dokumencie *j*.

Biorąc pod uwagę indeksy wektorów dla dwóch dokumentów, możliwe jest obliczenie współczynnika podobieństwa między nimi, $sim(d_i, d_j)$, który odzwierciedla stopień podobieństwa i wagi odpowiednich termów. Zamiast identyfikacji każdego dokumentu przez kompletny wektor pochodny na układzie współrzędnych, względna odległość między wektorami jest zachowana poprzez normalizację wszystkich długości wektorowych do jednej, uwzględniając rzut wektorów w zakresie przestrzeni przez nie reprezentowanej. W tym przypadku, każdy dokument może być przedstawiony przez pojedynczy punkt, którego pozycja jest określona przez obszar, w którym odpowiedni wektor dokumentu styka się z obwiednią przestrzeni. Dokumenty z dwoma podobnymi indeksami termu są wtedy reprezentowane przez punkty ulokowane bardzo blisko siebie w przestrzeni i w zasadzie odległość między dwoma dokumentami w przestrzeni jest odwrotnie skorelowana z podobieństwem odpowiadających im wektorów [14].

W modelu wektorowym podobieństwo dwóch wektorów określa się zazwyczaj poprzez miary związane z odległością tych wektorów w przestrzeni. Najczęściej stosowana jest miara kosinusowa [4, 21]. Gdy wektory są znormalizowane ta miara odpowiada iloczynowi skalarnemu $q \times d_j$ zdefiniowanemu jako:

$$q \times d_j = \sum_{i=1}^{n} q^i d_j^i \tag{2}$$

gdzie wektor *q* reprezentuje to czego szukamy – zapytanie, zaś d_j reprezentuje wektory dokumentów, $j=1,2,\ldots\ldots,n$. Jej podstawową zaletą jest to, iż mierząc kąt ignorowana jest długość dokumentu [10].

Istnieje alternatywa dla miary kosinusowej, wykorzystująca wszystkie termy, jakie występują w dokumentach, ale używa prostszej reprezentacji binarnej. Jej idea polega na rozważeniu tylko niezerowych współrzędnych binarnych wektorów. Obliczany jest współczynnik Jaccarda definiowany jako procent niezerowych współrzędnych, różnych dla obu wektorów [8]. Jeśli przyjmiemy, że *T(d)* oznacza zbiór termów występujących w dokumencie *d*, to podobieństwo między dwoma binarnymi wektorami dokumentów $sim(d_1,d_2)$ jest definiowane jako:

$$sim(d_1, d_2) = \frac{\left|T(d_1) \cap T(d_2)\right|}{\left|T(d_1) \cup T(d_2)\right|} \tag{3}$$

Równanie (3) służy do obliczenia podobieństwa między dokumentami wykorzystując alternatywny współczynnik Jaccarda zdefiniowany na zbiorach. Podobieństwo to osiąga wartość maksymalną, jeżeli dwa dokumenty są identyczne ($sim(d_1,d_2)=1$), a także jest symetryczne ($sim(d_1,d_2)=sim(d_2,d_1)$) [8].

W modelu wektorowym tracone są wszystkie informacje na temat struktury dokumentów: podział na zdania, nagłówki itp. Pomijane są informacje o kolejności słów i związków między nimi – dokument jest traktowany jako worek słów (ang. *bag of words*). Z drugiej strony, w modelu tym możliwe jest zastosowanie wielu istniejących technik sztucznej inteligencji, czy metod statystycznych. Także operacje takie jak, liczenie odległości, czy ważenie słów na wektorach są łatwiejsze i bardziej wydajne obliczeniowo od innych konkurencyjnych modeli np. grafowego [14].

1.2. Macierz częstości

Dokumenty w kolekcji są dzielone na termy. Zostają one oznaczane numerami od *1* do *t*, gdzie *t* oznacza liczbę wszystkich termów w dokumencie. Każdy dokument w kolekcji jest wtedy reprezentowany przez *t*-wymiarowy wektor częstości termu, zaś kolekcja *d* dokumentów jest reprezentowana przez $t \times d$ macierz częstości. W bardzo dużych kolekcjach dokumentów tekstowych wielkość zasobu słów (termów) jest rzędu dziesiątek tysięcy, powodując tworzenie niezmiernie rzadkich macierzy częstości. Taka „surowa" macierz częstości jest wtedy poddawana różnym transformacjom. Ma to na celu redukcję wymiaru macierzy, wynikającą z przekonania, że różnica między występowaniem termu dziesięć razy w porównaniu do jedenastu razy nie jest znacząca, tak samo jak wystąpieniem termu raz a wcale.

O stopniu podobieństwa między dokumentami informuje bardziej fakt wystąpienia takich samych termów niż precyzyjnie określona liczba ich wystąpień. Dlatego też wyznaczone częstości wystąpień poddać należy transformacji służącej uwypukleniu cech wspólnych dokumentów. Do najczęściej stosowanych metod transformacji macierzy częstości należy zastosowanie odpowiedniej reprezentacji dokumentu tekstowego.

Reprezentacja binarna

W reprezentacji binarnej zapamiętywany jest sam fakt wystąpienia *i*-tego słowa w *j*-tym dokumencie, nie precyzuje się liczby wystąpień. Zastosowanie tej metody prowadzi do zastąpienia oryginalnej macierzy częstości macierzą wystąpień o elementach a_{ij} równych:

- jeden, jeśli *i*-te słowo występuje w *j*-tym dokumencie (co najmniej raz),
- zero, jeśli *i*-te słowo nie występuje w *j*-tym dokumencie.

Zaletą reprezentacji binarnej jest prostota realizacji, oraz fakt użycia mniejszej mocy obliczeniowej dla wyznaczenia wartości niektórych metryk dla wektorów zero-jedynkowych. Przyjęcie tego modelu nadawania wag ma jednak swoje konsekwencje. Reprezentacja ta nie uwzględnia faktu, iż niektóre termy niosą więcej informacji niż pozostałe. Na podstawie oznaczenia czy dane słowo występuje w dokumencie czy też nie, trudno jest wnioskować o tematyce dokumentu, a wręcz mylnie definiować jego tematykę. Nie jest również uwzględniana informacja, w ilu dokumentach dane słowo występuje [12]. Reprezentacja ta dobrze się sprawdza w przypadku klasyfikacji dokumentów oraz ich grupowania, nie jest jednak odpowiednia do wyszukiwania słów kluczowych, ponieważ nie umożliwia ich uszeregowania [8].

Reprezentacja TF-IDF

Istnieje wiele sposobów obliczania współrzędnych wektora odpowiadającego konkretnemu dokumentowi. Wszystkie podstawowe metody bazują na wykorzystaniu trzech współczynników, jakimi można opisać dokument oraz jego związek z innymi dokumentami. Należą do nich [9]:

- częstotliwość termu (ang. *term frequency*, *tf*) – jest ilorazem wystąpień *i*-tego termu do sumy wystąpień wszystkich termów w j-tym dokumencie – innymi słowy normalizacja TF skaluje *j*-ty wektor tak, by suma jego składowych dawała jeden.

$$tf_{i,j} = \frac{tc_{i,j}}{\sum_k tc_{k,j}} \tag{4}$$

- częstotliwość dokumentu (ang. *document frequency*, *df*) – zlicza liczbę wystąpień danego termu w kolekcji dokumentów. Po jej zastosowaniu otrzymujemy dla każdego słowa wyliczoną wartość uzależnioną od liczby jego wystąpień. Następnie odrzucane są słowa, które mają skrajne wartości, występują w bardzo dużej bądź też bardzo małej liczbie dokumentów. Ważnym czynniki jest także odwrotna częstotliwość dokumentu

(ang. *inverse document frequency, idf)* będąca miarą popularności danego termu w kolekcji dokumentów. Im częściej występuje dany term, tym mniejsza jest jego waga *idf*. Jest ona wyrażana za pomocą następującego równania:

$$idf_i = \log \frac{|D|}{1 + |\{d \in D : d_i \neq 0\}|} \tag{5}$$

gdzie licznik w ułamku jest liczbą dokumentów w kolekcji, w mianowniku zaś znajduje się liczba dokumentów zawierających *i*-tego terma [16].

– częstotliwość kolekcji (ang. *collection frequency*, cf) [8].

Kombinacja przedstawionych dwóch parametrów, częstotliwości termu i odwrotnej częstotliwości w dokumentach, oznaczana jako TF-IDF jest najchętniej stosowanym schematem ważenia termów [12]. Algorytm TF-IDF powstaje z iloczynu $tf_{i,j} \cdot idf_i$. Wysoki współczynnik TF-IDF otrzymuje się dla termów występujących licznie w dokumencie, ale nie występujących zbyt często w innych dokumentach. Takie właśnie cechy najlepiej pozwalają na wykrycie różnic i podobieństw między tekstami [1].

Miara TF-IDF wymaga nie tylko analizy jednego dokumentu, ale także korzystanie z charakterystyki całej kolekcji dokumentów. Dodatkowo należy pamiętać o konieczności wyznaczania nowych wartości wag TF-IDF dla wszystkich termów każdego dokumentu zarówno w chwili zmiany zawartości dokumentu, jak również zmiany kolekcji przez dodanie bądź usunięcie dokumentu. W przypadku dużych kolekcji liczba różnych termów, a co za tym idzie zajętość pamięci może być znaczna.

2. WYBRANE METODY IDENTYFIKACJI SŁÓW KLUCZOWYCH

Proces identyfikacji słów kluczowych można podzielić na kilka etapów:

– identyfikacja termów występujących w dokumentach;
– usuwanie słów popularnych – zastosowanie stop listy;
– zliczanie wystąpień termów (obliczanie tzw. *tf – term frequency*);
– obliczanie wag dla wszystkich termów;
– przypisanie każdemu dokumentowi przynależnych prostych termów, które mogą odgrywać rolę słów kluczowych dokumentu.

2.1. Metody bazujące na macierzy częstości

Tematyka większości dokumentów jest zazwyczaj wystarczająco dobrze określona przez niewielką liczbę słów kluczowych, pozostałe informacje są zbędne. Potrzebna jest więc funkcja wybierająca słowa najbardziej istotne dla zbioru dokumentów, dla której dziedzinę będą stanowić słowa, zaś wartości określać będą ich przydatność do dalszej analizy. Funkcje ważące (ang. *weighting functions*) spełniają to zadanie ulepszając macierz częstości występowania słów w dokumencie.

Ważenie jest procesem, który każdemu słowu w dokumencie przypisuje wagę wynikającą z częstości jego wystąpień w dokumencie [6]. Najprostszym sposobem ważenia macierzy jest przypisanie każdej współrzędnej wektora dokumentu d częstości występowania słowa (termu) t w dokumencie. Schemat ten jest określany jako *term frequency* i oznacza się go jako $tf_{t,d}$ [6].

Opisana powyżej operacja prowadzi do definicji wskaźnika istotności słowa w postaci:

$$WIS^A_{t,d} = tf_{t,d} \tag{6}$$

gdzie:

$WIS^A_{t,d}$ – wskaźnik istotności t-słowa w d -tym dokumencie oparty na częstości wystąpienia.

Ta prosta metoda ma poważną wadę – każde słowo w dokumencie jest uznawane za jednakowo ważne. Również należy zauważyć, że wartość wskaźnika jest uzależniona od długości dokumentu.

Chcąc wyeliminować wpływ długości dokumentu można dokonać przekształcenia równania (6) zastępując wszystkie dodatnie wartości $tf_{t,d}$ przez 1, zaś wartości zerowe pozostawiając niezmienione. Prowadzi to do wskaźnika istotności słowa w postaci:

$$WIS_{t,d}^{B} = \begin{cases} tf_{t,d} = 1 \\ tf_{t,d} = 0 \end{cases} \tag{7}$$

gdzie:

$WIS_{t,d}^{B}$ – wskaźnik istotności t-słowa w d-tym dokumencie oparty na jego wystąpieniu, równy jedności jeśli t-słowo występuje w d-tym dokumencie (jeden bądź więcej razy) oraz równy zero jedności jeśli t-słowo nie występuje w d-tym dokumencie.

Próbą realizacji potrzeby zróżnicowania znaczenia poszczególnych słów w dokumencie może być przeskalowanie wartości macierzy $tf_{t,d}$ przez częstotliwość kolekcji (ang. *collection frequency*) cf [6]. Jednakże praktyka badawcza pokazuje, że lepszym rozwiązaniem jest uwzględnienie liczby dokumentów w których dane słowo (term) występuje – częstotliwość dokumentu df_t. Wartości df_t są tym większe im słowo t występuje w większej liczbie dokumentów. W formule obliczeniowej stosuje się odwrotną częstotliwość dokumentu idf_t zdefiniowaną jako $1/idf_t$, która jest wysoka dla słów występujących rzadko, zaś niska dla często występujących słów. W wyniku połączenia opisanych wyżej dwóch wag otrzymuje się definicję jednego z najbardziej popularnych schematów ważenia dokumentów w dziedzinie wydobywania informacji TF-IDF [14, 6]. Odpowiednie równanie przyjmuje więc postać:

$$WIS_{t,d}^{C} = tf_{t,d} \cdot idf_t = tf_{t,d} \cdot \log_2(N/df_t) \tag{8}$$

gdzie:

N – łączna liczba dokumentów,

$WIS_{t,d}^{C}$ – wskaźnik istotności t-słowa w d-tym dokumencie oparty na reprezentacji TF-IDF.

Zastosowanie równania (8) prowadzi do uzyskania wskaźników istotności słowa, które przyjmują:

- wartości maksymalne dla termów występujących często w małej liczbie dokumentów;
- wartości niskie dla termów występujących rzadko w małej liczbie dokumentów, lub występujących w dużej liczbie dokumentów, przez co termy te mają małą siłę rozróżniającą dokumenty;
- wartości minimalne dla termów pojawiających się w (prawie) wszystkich dokumentach.

3. BADANIA I WYNIKI

W artykule podjęto próbę zastosowania wybranych metod identyfikacji słów kluczowych do kolekcji danych tekstowych obejmujących streszczenia artykułów naukowych z zakresu zastosowania metod statystycznych w nauce, zarządzaniu i logistyce. Pierwotnym założeniem autora była analiza streszczeń artykułów z konferencji TransComp 2012, lecz ze względu na brak dostępnych słów kluczowych podanych przez autorów abstraktów założenie nie zostało zrealizowane. Nie istniałaby możliwość oceny i porównania wyników.

Analiza podanego zbioru dokumentów tekstowych została podzielona dodatkowo na następujące podtematy – zestawienia wyników:

– streszczenia artykułów naukowych całościowe – w każdym pojedynczym pliku jeden abstrakt;
– streszczenia artykułów naukowych z rozbiciem na zdania – w każdym pojedynczym pliku występuje tylko jedno zdanie abstraktu; stąd na jedno streszczenie może składać się kilka bądź kilkanaście dokumentów.

Do realizacji części empirycznej zastosowane zostały następujące metody badawcze:

– implementacja skryptu w języku Java dotyczącego przekształcenia słów polskojęzycznych do ich formy podstawowej;
– podział analizowanych kolekcji dokumentów wraz z ich rozbiciem na pliki składające się tylko z jednego zdania – do tego celu zostały wykorzystane możliwości jakie oferuje język R;
– analiza kolekcji danych tekstowych za pomocą języka R, a w szczególności pakietu tm.

Zestaw streszczeń artykułów zawiera także ręcznie przypisane przez autorów artykułów słowa kluczowe. Umożliwiają one określenie skuteczności zastosowania danej metody identyfikacji słów kluczowych.

3.1. Wyniki badań

Procedurę badawczą zapoczątkowało przekształcenie w analizowanych kolekcjach dokumentów tekstowych słów do ich formy podstawowej. Redukcja słów do ich formy podstawowej nie uwzględnia kontekstu użycia danego słowa, jednak uzyskane wyniki nie wpływają na znaczną utratę ich wartości informacyjnej.

W takcie badań metod identyfikacji słów kluczowych bazującej na macierzy częstości i dwóch jej przekształceniach: macierzy binarnej oraz ważonej logarytmicznej macierzy częstości wyznaczono trzy wartości wskaźników istotności słów w dwóch wersjach: nie uwzględniając stop-listy oraz z zastosowaniem stop-listy.

W celu określenia istotności słowa w całym zbiorze dokumentów wyznaczono dla poszczególnych słów sumę wskaźników cząstkowych obliczonych dla poszczególnych dokumentów. Przyjęto, że wyższa wartość wskaźnika świadczy o większym znaczeniu danego wyrazu. W trakcie obliczeń uwzględniono jedynie te wyrazy, które występują przynajmniej w dwóch dokumentach zbioru. Obliczenia zrealizowano w pakiecie R.

W tabeli (Tab. 1.) przedstawiono dziesięcioelementowe listy najistotniejszych wyrazów uzyskane za pomocą każdej z metod oraz wartość wskaźnika istotności.

Tab. 1. Lista uzyskanych najistotniejszych wyrazów z podziałem na metody bazujące na macierzy częstości dla analizowanego zbioru dokumentów.

Badanie nieuwzględniające stop-listy		
Analiza streszczeń artykułów naukowych		
$WIS_{t,d}^{A}$	$WIS_{t,d}^{B}$	$WIS_{t,d}^{C}$
być (232)	być (53)	model (110,444302)
siebie (118)	oraz (38)	nowotwór (105,189914)
oraz (106)	siebie (37)	wskaźnik (91,56048)
dany (91)	który (35)	gospodarstwo (88,78878)
który (78)	dany (32)	badanie (83,342537)
analiza (73)	ten (31)	rozwój (83,342537)
dla (73)	analiza (30)	strategia (79,566927)
statystyczny (72)	dla (30)	region (77,324125)
cel (68)	cel (28)	gospodarczy (75,012182)
praca (68)	metoda (28)	badań (74,027973)
Badanie uwzględniające stop-listy		
Analiza streszczeń artykułów naukowych		
$WIS_{t,d}^{A}$	$WIS_{t,d}^{B}$	$WIS_{t,d}^{C}$
analiza (73)	analiza (30)	model (110,444302)

statystyczny (72)	praca (28)	nowotwór (105,189914)
praca (68)	statystyczny (28)	gospodarstwo (88,78878)
model (62)	wykorzystanie (26)	badanie (83,342537)
badanie (60)	informacja (23)	strategia (79,566927)
statystyk (49)	badanie (21)	region (77,324125)
informacja (47)	statystyk (21)	gospodarczy (75,012182)
polski (44)	społeczny (19)	domowy (73,120172)
wykorzystanie (41)	główny (18)	polski (70,903127)
gospodarczy (38)	polski (18)	statystyczny (70,128345)

Źródło: opracowanie własne

W przypadku analizy streszczeń artykułów naukowych dokonano porównania wyników uzyskanych za pomocą badanych algorytmów z rzeczywistą listą słów kluczowych, która została określona przez autorów abstraktów. W tym celu:

- Przeprowadzona została analiza plików zawierających słowa kluczowe;
- Uzyskana lista słów kluczowych wraz z obliczonymi wskaźnikami istotności zostaje ograniczona do 25 początkowych wyrazów stanowiących najważniejsze słowa kluczowe. Podobnie dla analizowanego zbioru streszczeń artykułów naukowych, dla którego obliczono wskaźnik $WIS_{t,d}^{A}$ lista słów zostaje zredukowana do 25.
- Mając dwa wyżej wymienione zbiory słów dla określenia ich prawdopodobieństwa obliczony zostaje indeks Jaccarda.
- Wartość obliczonego indeksu Jaccarda przyjęta zostaje, jako miara poprawności.

Ocenę metod bazujących na macierzy częstości przedstawia poniższa tabela (Tab. 2).

Tab. 2. Ocena metod wyznaczania wskaźników istotności słów z uwzględnieniem kryterium bazującym na macierzy częstości.

Kryterium	Ocena
Tryb tworzenia modelu bazowego	Modelem bazowym dla analizowanej grupy wskaźników istotności słów jest model przestrzeni wektorowej konstruowany na podstawie macierzy częstości. W badaniach wykorzystano dwie wersje macierzy częstości – pierwsza tworzona była bez uwzględnienia stop-listy, w drugiej uwzględniono stop-listę.
Zakres informacji uwzględnianej w trakcie oceny istotności słów	Wskaźnik $WIS_{t,d}^{A}$ i $WIS_{t,d}^{B}$ może zostać wyznaczony niezależnie dla poszczególnych dokumentów. Wskaźnik $WIS_{t,d}^{C}$ można wyznaczyć jedynie na podstawie całego zbioru dokumentów (jego obliczenie dla pojedynczego dokumentu wymaga znajomości odwrotnej częstości dokumentowej, która szacowana jest na podstawie zbioru dokumentów).
Poprawność działania	Badania pokazały, że podejście bez stosowania stop-listy nie pozwoliło na uzyskanie poprawnych rozwiązań (wiele wyrazów zidentyfikowanych przez metodę jako istotne nie posiada dużej wartości informacyjnej). W przypadku analizy streszczeń artykułów naukowych dokonano porównania wyników uzyskanych za pomocą badanych algorytmów z rzeczywistą listą słów kluczowych, która została określona przez autorów abstraktów. Miarą poprawności jest obliczona wartość indeksu Jaccarda, który dla badania nieuwzględniającego stop-listy wynosi 0,2195122, zaś dla badania uwzględniającego stop-listę 0,3157895. Wyższa wartość współczynnika Jaccarda wskazuje na większe prawdopodobieństwo wystąpienia uzyskanych słów w zbiorze słów kluczowych charakteryzujących całą kolekcję.

Źródło: opracowanie własne

PODSUMOWANIE

Uogólniając wyniki badań można sformułować następujące wnioski w zakresie skuteczności omówionych metod identyfikacji słów i fraz kluczowych dla zbioru streszczeń artykułów naukowych z dziedziny logistyka:

1. W przypadku tworzenia macierzy częstości w modelu przestrzeni wektorowej zdecydowanie lepsze wyniki skuteczności zastosowanych metod osiągnięte zostały po zastosowaniu stop-listy dla streszczeń artykułów naukowych.
2. W przypadku badania metod bazujących na podstawowej macierzy częstości oraz jej reprezentacji binarnej i uzyskujemy bardzo zbliżone i poprawne wyniki dla streszczeń artykułów naukowych. Natomiast w przypadku zastosowania modyfikacji macierzy uwzględniającej TF-IDF uzyskany zbiór słów kluczowych jest inny i wskazuje na lepsze całościowe odzwierciedlenie treści badanych abstraktów.
3. W przypadku analizy streszczeń artykułów naukowych dokonano porównania wyników uzyskanych za pomocą badanych algorytmów z rzeczywistą listą słów kluczowych, która została określona przez autorów abstraktów. Dokonując wyznaczenia indeksu Jaccarda, jako miary poprawności najlepsze wyniki uzyskano dla zbioru uwzględniającego stop-listę.

Reasumując dotychczasowe rozważania należy zauważyć, iż w celu określenia skuteczności analizowanych metod algebraicznych opartych na modelu przestrzeni wektorowej należy rozszerzyć badanie na szerszy wachlarz istniejących metod wykorzystywanych do identyfikacji słów kluczowych: oparte dekompozycji SVD, metodzie LDA czy chmurze słów. Dla zbioru analizowanych dokumentów, w których występuje język naukowy specyficzny dla dziedziny logistyki wykorzystując analizowane metody, wyniki są poprawne. Należy oczekiwać zdecydowanie lepszych wyników proponując rozwiązania pozwalające na identyfikację słów i fraz kluczowych przy wykorzystaniu wiedzy dziedzinowej opisanej w postaci sieci semantycznej lub innej metody reprezentacji wiedzy.

Zamiarem autora jest w kolejnych artykułach wchodzących w skład rozpoczętego cyklu publikacji dokonanie analizy i przeprowadzenie badań innych wykorzystywanych metod algebraicznych do identyfikacji słów kluczowych w dokumentach tekstowych.

BIBLIOGRAFIA

1. Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, ACM Press, Adison-Wesley, New York 1999.
2. Ball R., *The scientific information environment in the next millennium*, Library Management, 2000, vol. 21, no. 1, s. 10-12.
3. Fayyad U. M., Piatetsky-Shapiro G., Smyth P., *From data mining to knowledge discovery in databases*, AI Magazine, Vol 17, 1996, s. 37-54.
4. Konchady M., *Text Mining Application Programming*, Charles River Media, 2006.
5. Leopold E., Kindermann J., *Text categorization with support vector machines, how to represent texts in input space?*, Machine Learning, 46, 2002.
6. Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England 2008.
7. Manning C. D., Schütze H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge Mass 2001.
8. Markov Z., Larose D. T., *Eksploracja zasobów internetowych*, PWN, Warszawa 2009.
9. Raghan V. V., Wong S, K, M., *A critical analysis of the vector space model for information retrieval*, Journal of the American Society for Information Science, 37(5), 1986, s. 100-124.

10. Schenker A., Bunke H., Last M., Kandel A., *Graph-Theoretic Techniques for Web Content Mining*, World Scientific Publishing Co, Pte, Ltd., 2005.
11. Smyth P., Goodman R, M., *Rule introduction using information theory*, In Knowledge Discovery in Databases, 1991, s. 159-176.
12. Salton G., Buckley Ch., *Term-Weighting Approaches in Automatic Text Retrieval*, Information Processing & Management, 24 (5), 1988, s. 513–523.
13. Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw–Hill Book Co., New York 1983, s. 52–117.
14. Salton G., Wong A., Yang C. S., *A vector space model for automatic indexing*, Communications of the ACM, vol, 18, 1975, s. 613–620.
15. Turing A., *Computing machinery and intelligence*, Mind, nr 236, 1950, przedruk w The Mind's I, Penguin Books, 1981.
16. Xia F., Jicun T., Zhihui L., *A Text Categorization Method Based on Local Document Frequency*, Proceedings in Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD '09, 2009, s. 468-471.
17. Zhang T., Oles F., *Text categorization based on regularized linear classification methods*, Information Retreival 2001, no 1.

METHODS FOR AUTOMATIC KEYWORDS IDENTIFICATION IN TEXT

Abstract

Development of the information society and information technology entailed an a natural creation of automated systems supporting find and organize information. Too much information stored in text documents is extremely important to for automatic keywords identification. Article begins a series dedicated to the study of algebraic methods used to for keywords identification in scientific Polish texts.

Autorzy:
dr inż. **Anna Gładysz** – Politechnika Rzeszowska, Wydział Zarządzania, Zakład Informatyki w Zarządzaniu