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

GŁADYSZ Anna

WYKORZYSTANIE DEKOMPOZYCJI SVD DO AUTOMATYCZNEJ IDENTYFIKACJI SŁÓW KLUCZOWYCH W DOKUMENTACH

Streszczenie

Artykuł stanowi kontynuację cyklu badań związanych z wykorzystaniem algebraicznych metod identyfikacji słów kluczowych w dokumentach tekstowych. Jego celem jest teoretyczna analiza i empiryczna weryfikacja przydatności użycia metod identyfikacji słów kluczowych opartej na dekompozycji SVD w naukowych tekstach polskojęzycznych.

WSTĘP

Gwałtowny rozwój technologii obliczeniowej spowodował, że współczesny człowiek zalewany jest strumieniami danych. Nadmiar informacji zmusza odbiorcę do wzmożonego wysiłku wkładanego w ich odbiór i selekcje, dlatego kluczowego znaczenia nabiera umiejętność zarządzania informacją. Jedną z najpopularniejszych i najczęściej spotykanych form zapisu informacji stanowią dokumenty tekstowe. Mogą one podlegać eksploracji danych, zaś działalność związana z ich przetwarzaniem może być w dużym stopniu zautomatyzowana dzięki eksploracyjnej analizie tekstu określanej jako text mining. Text mining definiowany jest, jako odkrywanie i wykorzystanie wiedzy zawartej w zbiorze dokumentów [6]. Jedną z istotnych zalet eksploracyjnej analizy teksu jest uwzględnienie kontekstu, w którym pojawiają się szukane słowa i wyrażenia. Techniki eksploracyjnej analizy danych tekstowych znajdują coraz większe zastosowanie w ekonomii, logistyce, lingwistyce oraz innych dyscyplinach, w których występuje problem nadmiaru informacji zapisanej w języku naturalnym.

Artykuł stanowi kontynuację cyklu badań związanych z wykorzystaniem algebraicznych metod identyfikacji słów kluczowych w dokumentach tekstowych. Jego celem jest teoretyczna analiza i empiryczna weryfikacja przydatności użycia metod identyfikacji słów kluczowych opartej na dekompozycji SVD w naukowych tekstach polskojęzycznych z zakresu logistyka.

1. ROZKŁAD WEDŁUG WARTOŚCI OSOBLIWYCH (SVD)

Istotną wadą modelu opierającego się na różnych modyfikacjach macierzy częstości jest nieuwzględnienie żadnej semantycznej zależności pomiędzy poszczególnymi słowami (termami). Rozwiązaniem tego problemu jest odpowiednie przekształcenie macierzy częstości do postaci, która uwidacznia tzw. ukryte znaczenie semantyczne (ang. *latent semantic*). Analiza przeprowadzana na tak zmodyfikowanej strukturze danych zwana jest w literaturze analizą LSA (ang. *latent sematic analysis*) lub indeksowaniem LSI (ang. *latent semantic indexing*) [1, 3].

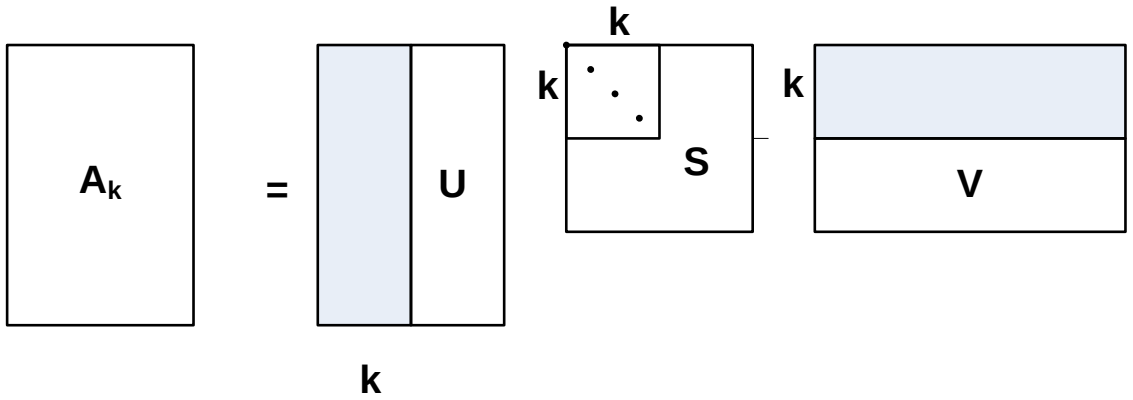
Narzędziem pozwalającym na dokonanie postulowanego przekształcenia jest dekompozycja macierzy częstości według wartości osobliwych. Pierwszą pracą, w której zastosowano to podejście jest [2].

Przekształcenie SVD polega na obliczeniu rozkładu macierzy częstości A o wymiarach $m \times n$ (można przyjąć założenie, że $m \geq n$) w postaci trzech macierzy określonych równaniem:

$$A = USV^T \tag{1}$$

gdzie $U_{m \times m}$, $V_{n \times n}$ są macierzami ortonormalnymi, zaś $S_{k \times k}$ jest macierzą diagonalną, której elementy zwane są wartościami osobliwymi (ang. *singular values*). Elementy te są umieszczone na diagonalnej macierzy w kolejności malejącej. Kolumny macierzy U oraz V są zwane wektorami osobliwymi (ang. *singular vectors*). Kolumny macierzy U oraz V tworzą również nowe (ortogonalne) bazy dla przestrzeni kolumn/wierszy macierzy A [11, s.386].

Istotnym elementem przedstawionego powyżej modelu jest malejące uporządkowanie elementów diagonalnej S. Uwzględniając fakt, że kolejne wartości osobliwe reprezentują znaczenie kolejnych wymiarów nowej przestrzeni, można sformułować spostrzeżenie, ze każdy kolejny wymiar ma mniejsze znaczenie informacyjne niż wymiary go poprzedzające. Pozwala to na dokonanie aproksymacji macierzy A na podstawie k początkowych wymiarów zidentyfikowanych w trakcie analizy. Ideę takiego postepowania przedstawia Rys. 1.



Rys. 1. Interpretacja graficzna dekompozycji SVD.
Źródło: opracowanie własne na podstawie [2]

Interpretacja powyższego opisu w kontekście macierzy częstości dokumentów znajduje się w tabeli Tab. 1.

Tab. 1. Interpretacja komponentów dekompozycji SVD w kontekście analizy dokumentów tekstowych.

A_k	najlepsza aproksymacja rzędu k macierzy A	m	liczba termów
U	macierz wektorów termów (słów)	n	liczba dokumentów
S	macierz wartości osobliwych	k	liczba czynników
V	macierz wektorów dokumentów	r	rząd macierzy A

Źródło: opracowanie własne

Wektorowy model dokumentu wykorzystywany jest w technice zwanej LSA (ang. *Latent Semantic Analysis*) opatentowanej w roku 1990 przez Deerwester'a, Dumas'a, Furnas'a i Landauer'a [2]. LSA zakłada, że zbiór dokumentów zawiera treści dotyczące znanej liczby tematów. Mimo, iż ilość tematów grup jest znana (założona z góry) dla konkretnego wykonania algorytmu, to cechy klasyfikujące poszczególne tematy są ukryte (nieznane). Celem metody LSA jest znalezienie dokumentów, które nie koniecznie maja ten sam zadany

zbiór słów, ale są na ten sam temat [9]. Punktem wyjścia modelu jest reprezentacja traktująca dokument jako zbiór słów. W przypadku metody LSA jest to macierz częstości. Macierz ta jest poddawana analizie SVD. Analiza składników głównych pojawiła się na początku XX wieku za sprawą Karla Pearsona [10, s. 559–572], natomiast swój rozwój zawdzięcza pracy Hotellinga [7, s. 417–444]. Podstawowe znaczenie rozkładu SVD dla metod numerycznych algebry liniowej stało się jasne dopiero w roku 1965 po opublikowaniu przez Goluba i Kahana konstruktywnej, numerycznie stabilnej metody jego wyznaczania [4, s. 205-224]. Rozkład SVD polega na dekompozycji macierzy na iloczyn trzech specyficznych macierzy w postaci $A = USV^T$, gdzie macierz U reprezentuje wyrazy w przestrzeni wyznaczone przez składowe, macierz V reprezentuje dokumenty w przestrzeni wyznaczone przez składowe, zaś macierz S jest diagonalna i zawiera znaczenie kolejnych składowych. Wyznaczone składowe odpowiadają tematom występującym w dokumentach. Zastosowanie SVD powoduje, że dokumenty i słowa są przedstawiane w przestrzeni tematów – co pozwala na analizę związków pomiędzy dokumentami, pomiędzy słowami oraz pomiędzy dokumentami i słowami.

Jednym z zastosowań dekompozycji SVD jest lepsze uwarunkowanie rozwiązania układu normalnego. SVD jest też dość często używane do rozwiązywania układów nadokreślonych [12]. W analizie danych tekstowych jest ona wykorzystywana do redukcji wymiaru macierzy częstości. Zastosowanie techniki SVD polepsza w pewnych przypadkach efektywność wyszukiwania w usystematyzowanych tekstach, dzięki możliwości dopasowania dokumentów i zapytań, które nie mają żadnych wspólnych wyrażeń [5]. Jej zastosowanie w grupowaniu dokumentów jest wciąż tematem badań [8]. Metoda ta jest dogodnym narzędziem przetwarzania informacji w postaci listy wyrazów, zaś jej wyniki stanowią punkt wyjścia dalszych obliczeń.

1.1. Metoda identyfikacji słów kluczowych oparta na dekompozycji SVD

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.

Opisana powyżej analiza LSA może stanowić podstawę do sformułowania kolejnej propozycji wskaźników istotności słów (*WIS*).

Punktem wyjścia jest dekompozycja wariancji (przyjmowanej jako miara wartości informacyjnej) wyznaczonych składowych (czynników) według słów uwzględnionych w trakcie analizy. Przyjmując, że w macierzy częstości słowom odpowiadają wiersze, zaś dokumentom kolumny, macierz wariancji-kowariancji dla aproksymacji macierzy częstości uwzględniających k – składowych przyjmuje postać:

$$A_k A_k^T = USV^T VSU^T = USSU^T \tag{2}$$

Za wskaźnik istotności poszczególnych słów należy uznać wartości diagonalne wyznaczonej powyżej macierzy:

$$WIS_t^D = diag(USSU^T) \tag{3}$$

2. BADANIA I WYNIKI

Do badania użyto zbioru dokumentów tekstowych zawierających streszczenia artykułów naukowych dotyczących wykorzystania metod statystycznych w nauce, zarządzaniu i

logistyce. 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. Należy jednak zauważyć, że czasami wybrane przez autora słowa kluczowe wynikają z analizy całej treści artykułu, nie zaś jego streszczenia. Wiąże się to z faktem występowania w słowach kluczowych fraz które nie są ani razu użyte w streszczeniu danego artykułu.

Przy ocenie wybranych metod identyfikacji słów kluczowych poddano analizie następujące aspekty ich działania lub wykorzystania:

1. Tryb tworzenia modelu.
2. Zakres informacji wykorzystywanej w trakcie uczenia modelu - biorąc pod uwagę ten aspekt wyróżnia się:
   - metody korpusowe – wykorzystujące przy analizie pojedynczego dokumentu wiedzę pochodzącą z całego zbioru dokumentów;
   - metody dokumentowe – przy analizie jednego dokumentu operują jedynie na tym tekście i nie wykorzystują żadnych informacji dotyczących innych dokumentów wchodzących w skład zbioru.
3. Poprawność działania – mierzona stopniem zgodności pomiędzy zbiorami słów kluczowych zidentyfikowanych przez człowieka oraz wykrytych przez algorytm. Jako miernik może zostać użyty współczynnik Jaccarda.

2.1. Wyniki badań

Procedurę badawczą zapoczątkowało przekształcenie w analizowanych kolekcjach dokumentów tekstowych słów do ich formy podstawowej. Program został napisany w języku Java z wykorzystaniem biblioteki *morfologik-stemming* w wersji 1.5.4. Redukcja słów do ich formy podstawowej nie uwzględnia kontekstu użycia danego słowa. Należy jednak stwierdzić, że w analizowanych kolekcjach dokumentów uzyskane wyniki redukcji słów do form podstawowych nie wpływają na znaczną utratę ich wartości informacyjnej.

Punktem wyjścia do badań była macierz częstości utworzona na podstawie całego analizowanego zbioru dokumentów z uwzględnieniem utworzonej listy słów istotnych i zastosowaniem stop-listy oraz dwie jej modyfikacje: macierz binarna oraz ważona logarytmiczna macierz częstości.

Analizowane kolekcje dokumentów tekstowych zostały poddane analizie na dwa sposoby:

- jako całe obrobione teksty (składające się z różnej większej liczby zdań) poddane redukcji do rdzenia dla języka polskiego;
- jako teksty podzielone na pojedyncze zdania, przy czym każde pojedyncze zdanie stanowi odrębny dokument tekstowy – również przepuszczone przez program dokonujący redukcji do rdzenia słów w języku polskim.

Podział tekstu umieszczonego w jednym pliku na zdania z których każde zostało zapisane w osobnym dokumencie tekstowym został zaimplementowany w języku R. W tym celu została wykorzystana biblioteka *stringr*, zaś separator końca zdania i początku następnego został ustawiony na kropkę.

W trakcie badań wyznaczono wskaźnik WIS_t^D (równanie (3)) w sześciu różnych wersjach:

- WIS_t^{D1} - na podstawie macierzy częstości;
- WIS_t^{D2} - na podstawie binarnej macierzy częstości;
- WIS_t^{D3} - na podstawie ważonej macierzy częstości TFIDF;
- WIS_t^{D4}- na podstawie macierzy częstości, przy czym każde zdanie w dokumencie traktowano jako oddzielny dokument;
- WIS_t^{D5}- na podstawie binarnej macierzy częstości, przy czym każde zdanie w dokumencie traktowano jako oddzielny dokument;
- WIS_t^{D6} - na podstawie ważonej macierzy częstości TFIDF, przy czym każde zdanie w dokumencie traktowano jako oddzielny dokument.

Przyjęto, że wyższa wartość wskaźnika świadczy o większym znaczeniu danego słowa.

W poniższej tabeli (Tab. 2) przedstawiono listy najistotniejszych wyrazów uzyskane za pomocą metody opartej na dekompozycji SVD wraz z wartościami wskaźnika istotności.

Tab. 2. Istotność słów uzyskana metodą opartą na dekompozycji SVD analizowanych kolekcji dokumentów.

Analiza streszczeń artykułów naukowych – bez rozbicia na zdania					
WIS_t^{D1}		WIS_t^{D2}		WIS_t^{D3}	
wyraz	**wskaźnik istotności**	**wyraz**	**wskaźnik istotności**	**wyraz**	**wskaźnik istotności**
model	598,7920	analiza	28,4321	nowotwór	6179,0900
statystyczny	427,4486	statystyczny	26,9104	wiejski	5642,7620
praca	327,1641	praca	26,5437	grunt	2704,5190
nowotwór	264,9413	wykorzystanie	24,2200	macierz	1932,3010
analiza	251,8711	informacja	21,5106	model	1837,2030
badanie	249,8325	statystyk	18,6634	głębia	1587,8390
obszar	233,9668	badanie	18,3111	starość	1587,1670
statystyk	223,4455	społeczny	16,5474	strategia	1568,3690
gospodarstwo	184,1032	problem	14,9299	dobrobyt	1563,4800
wiejski	164,5830	model	14,1024	złośliwy	1455,2810
Analiza streszczeń artykułów naukowych – z rozbiciem na zdania					
WIS_t^{D4}		WIS_t^{D5}		WIS_t^{D6}	
wyraz	**wskaźnik istotności**	**wyraz**	**wskaźnik istotności**	**wyraz**	**wskaźnik istotności**
badanie	107,7483	analiza	69,8671	badanie	1755,4340
model	101,8506	statystyczny	68,8825	model	1666,7850
analiza	78,8415	praca	62,8533	gospodarstwo	1269,9440
statystyczny	77,8547	badanie	48,7302	domowy	1126,7130
praca	77,7592	model	48,7104	informacja	1043,3080
informacja	58,6739	statystyk	44,6075	praca	1033,9290
statystyk	56,6250	informacja	41,7839	statystyk	967,1090
gospodarstwo	52,7059	wykorzystanie	40,7058	analiza	955,9668
gospodarczy	46,6344	społeczny	33,5657	statystyczny	955,5713
domowy	42,5706	gospodarczy	32,3307	gospodarczy	939,3007

Źródło: opracowanie własne

W przypadku analizy streszczeń artykułów naukowych dokonano porównania wyników uzyskanych na całościowej kolekcji (bez rozbicia na zdania), oraz kolekcji z rozbiciem na zdania z rzeczywistą listą słów kluczowych, która została określona przez autorów abstraktów. W tym celu:

- Przeprowadzona została analiza plików zawierających słowa kluczowe – słowa kluczowe określone przez autora dla każdego streszczenia zostały zapisane w osobnych plikach tekstowych. Słowa kluczowe zostały przekształcone do postaci podstawowej, następnie została utworzona macierz częstości, obliczone zostały wskaźniki WIS_t^{D1} oraz WIS_t^{D5} (oparte na częstotliwościach występowania), zsumowane dla wszystkich dokumentów, zaś na sam koniec zostały uszeregowane malejąco słowa kluczowe według wyliczonego wskaźnika istotności.
- Uzyskane w ten sposób listy słów kluczowych wraz z obliczonymi wskaźnikami istotności zostały ograniczone do 25 początkowych wyrazów stanowiących najważniejsze słowa kluczowe.
- Podobnie dla analizowanego zbioru streszczeń artykułów naukowych, dla którego obliczono wskaźniki WIS_t^{D1} oraz WIS_t^{D5} listy słów zostały zredukowana do 25.
- Mając wyżej wymienione zbiory słów dla określenia ich prawdopodobieństwa obliczony został indeks Jaccarda. Aby obliczenia były poprawne zostały odpowiednio przekształcone badane obiekty w celu zdefiniowania mocy części wspólnej i sumy zbiorów. Aby porównać teksty w języku naturalnym dokonane zostało przekształcenie tekstu na postać zbioru, którego elementami są poszczególne słowa.
- Wartość obliczonego indeksu Jaccarda przyjęta została, jako miara poprawności. Założono, że podobieństwo dwóch tekstów w języku naturalnym zależy głównie od jednoczesnego występowania pewnych słów w obu tekstach.

Wyniki analizy okrojonych zbiorów słów powstałych bez rozbijania streszczeń na pojedyncze zdania dla artykułów naukowych wraz z wyliczonym indeksem Jaccarda zobrazowane zostały w tabeli Tab. 3.

Tab. 3. Obliczanie indeksu Jaccarda dla 25 słów z analizowanych zbiorów streszczeń artykułów naukowych (bez rozbicia na zdania) po dekompozycji SVD.

	Moc części wspólnej zbiorów	**Moc sumy zbiorów**	**Indeks Jaccarda**
	$\lvert T(d_1) \cap T(d_2) \rvert$	$\lvert T(d_1) \cup T(d_2) \rvert$	$\frac{\lvert T(d_1) \cap T(d_2) \rvert}{\lvert T(d_1) \cup T(d_2) \rvert}$
Wartości (liczba słów)	11	39	0,2820513
Wyrazy występujące w obu zbiorach			
"model", "statystyczny", "praca", "nowotwór", "analiza", "badanie", "obszar", "gospodarczy", "informacja", "system", "rynek"			

Źródło: opracowanie własne

Wyniki analizy okrojonych zbiorów słów powstałych z rozbicia streszczeń na pojedyncze zdania dla artykułów naukowych wraz z wyliczonym indeksem Jaccarda zobrazowane zostały w tabeli Tab. 4.

Tab. 4. Obliczanie indeksu Jaccarda dla 25 słów z analizowanych zbiorów streszczeń artykułów naukowych (z rozbiciem na zdania) po dekompozycji SVD.

	Moc części wspólnej zbiorów	**Moc sumy zbiorów**	**Indeks Jaccarda**
	$\lvert T(d_1) \cap T(d_2) \rvert$	$\lvert T(d_1) \cup T(d_2) \rvert$	$\frac{\lvert T(d_1) \cap T(d_2) \rvert}{\lvert T(d_1) \cup T(d_2) \rvert}$
Wartości (liczba słów)	13	37	0,3513514
Wyrazy występujące w obu zbiorach			
"badanie", "model", "analiza", "statystyczny", "praca", "informacja", "gospodarczy", "obszar", "społeczny", "rozkład", "system", "rynek", "nowotwór"			

Źródło: opracowanie własne

2.2. Ocena metody bazującej na dekompozycji SVD

Ocenę badanej grupy metod wyznaczania wskaźników istotności słów uzyskanych metodą opartą na dekompozycji SVD zgodnie z przyjętymi kryteriami przedstawia tabela Tab. 5.

Tab. 5. Ocena metod wyznaczania wskaźników istotności słów z uwzględnieniem metody bazującej na dekompozycji SVD.

Kryterium	**Ocena**
Tryb tworzenia modelu bazowego	Modelem bazowym jest dekompozycja macierzy częstości (w formie podstawowej, binarnej lub ważonej) względem wartości osobliwych.
Zakres informacji uwzględnianej w trakcie oceny istotności słów	Wskaźniki istotności słów mogą zostać wyznaczone wyłącznie na podstawie analizy pełnego zbioru dokumentów.
Poprawność działania	Badania pokazały, że uzyskano lepsze wyniki oparte na dekompozycji SVD przy uwzględnieniu analizowanego zbioru z podziałem na pojedyncze zdania dla streszczeń artykułów naukowych. 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 rozbicia streszczeń artykułów na pojedyncze zdania zapisane w oddzielnych plikach wynosi 0,2820513, zaś dla badania uwzględniającego podział na zdania 0,3513514. Wyższa wartość współczynnika Jaccarda wskazuje na większe prawdopodobieństwo wystąpienia uzyskanych słów za pomocą dekompozycji SVD w zbiorze streszczeń artykułów naukowych z rozbiciem na zdania.

Źródło: opracowanie własne

PODSUMOWANIE

Uogólniając wyniki badań można sformułować następujące wnioski w zakresie skuteczności omówionej metody identyfikacji słów kluczowych dla zbioru streszczeń artykułów naukowych z dziedziny logistyka:

1. Zastosowanie stop-listy oraz redukcji do rdzenia pozwala w istotny sposób zmniejszyć liczbę wymiarów macierzy częstości.
2. 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 dla metody wykorzystującej dekompozycję SVD uzyskano dla zbioru uwzględniającego rozbicie na pojedyncze zdania (każde zdanie streszczenia w osobnym dokumencie tekstowym). Należy wskazać także na stosunkowo dużą różnicę pomiędzy wyliczonymi wskaźnikami indeksu Jaccarda dla zbioru uwzględniającego rozbicie na zdania w stosunku do tego, który nie opiera się na podziale.

Reasumując dotychczasowe rozważania i biorąc pod uwagę wyniki uzyskane w oparciu o metody bazujące na macierzy częstości i jej modyfikacjach (opisane we wcześniejszym artykule), należy wskazać wyższość metody opartej na dekompozycji SVD. Wynika ona chociażby z obliczonej wartości indeksu Jaccarda, która jest największa dla analizowanej metody. Jednak 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 o metody oparte na LDA oraz chmurze słów. Wyniki uzyskane na podstawie analizy przebadanych metod są zadowalające. Jednak należy oczekiwać lepszych 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, w szczególności metody uwzględniającej LDA.

BIBLIOGRAFIA

1. Berry M. W., Dumais S. T., O'Brien G. W., *Using linear algebra for intelligent information retrieval*, SIAM Review, 37(4), 1995, s. 573-595.
2. Deerwester S., Dumais S. T., Furnas G., Landauer T. K, Harshman R., *Indexing by Latent Semantic Analysisis*, Journal of the American Society for Information Science, 41(6), 1990.
3. Furnas G. W., Deerwester S., Dumais S. T., Landauer T. K., Harshman R. A., Streeter L. A., Lochbaum K. E., *Information retrieval using a singular value decomposition model of latent semantic structure*, In Proceedings of the ACM SIGIR Conference, ACM, New York, 1998, s. 465-480.
4. Golub G., Kahan V., *Calculating the singular values and pseudoinverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis 2, 1965, s. 205-224.
5. Hand D., Mannila H., Smyth P., *Eksploracja danych*, Wydawnictwo Naukowo-Techniczne, Warszawa 2005.
6. Hearst M., *Untangling Text Data Mining*, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26 1999.
7. Hotteling H., *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, nr 24, 1933, s. 417–444.
8. Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England 2008.
9. Osińska V., *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych*, EBIB Nr76, 2006.

10. Pearson K., *On lines and planes of closest fit to system of points in space*, Philosophical Magazine, nr 2, 1901, s. 559–572.
11. Walesiak M., Gatnar E., *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa 2009.
12. Walker D., *Full and banded matrix algorithms*, Parallel Computing Works, 1996.

USAGE DECOMPOSITION SVD FOR AUTOMATIC KEYWORDS IDENTIFICATION IN TEXT DOCUMENTS

Abstract

The article is a continuation of the cycle of studies related to the use of algebraic methods for keywords identification in text documents. Its purpose is to theoretical analysis and empirical verification of the suitability of the use of methods for keywords identification based on SVD decomposition of scientific in Polish texts.

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