PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Katarzyna Wójcik

Uniwersytet Ekonomiczny w Krakowie

ANALIZA PORÓWNAWCZA MIAR PODOBIEŃSTWA TEKSTÓW

Streszczenie: Zasadniczym celem niniejszej pracy jest próba oceny przydatności znanych z literatury miar podobieństwa tekstów. W kolejnych punktach artykułu przedstawiono najpierw wybrane miary podobieństwa, oparte ma macierzy częstości, wykorzystane do porównania dokumentów, a następnie dokumenty tekstowe, które były porównywane w badaniu. W dalszej części zaprezentowano wyniki przeprowadzonej analizy symulacyjnej opisanych wcześniej miar. Na tej podstawie sformułowano wnioski dotyczące zbieżności między analizowanymi miarami a rzeczywistym podobieństwem pomiędzy dokumentami.

Słowa kluczowe: miara, podobieństwo, macierz częstości, *text mining.*

1. Wstęp

Jednym z najważniejszych problemów pojawiających się przy eksploracyjnej analizie danych tekstowych jest wybór sposobu wyrażenia podobieństwa pomiędzy tekstami. W literaturze prezentowane są dwa zasadnicze podejścia do rozwiązania tego problemu.

Pierwsze z nich bazuje na reprezentacji częstotliwościowej dokumentów (macierzy zawierającej informacje o liczbie wystąpień poszczególnych słów w dokumencie). Główną zaletą takiego podejścia jest prostota obliczeń oraz brak konieczności uwzględniania wiedzy lingwistycznej lub dziedzinowej. Wadą jest natomiast brak możliwości uwzględnienia związków pomiędzy słowami. Szczególną i jednocześnie często stosowaną podgrupą są miary wykorzystujące zredukowaną wersję macierzy częstości [Lula 2005].

Drugie podejście do obliczania podobieństwa między tekstami zakłada wykorzystanie wiedzy dziedzinowej. Jej uwzględnienie zwykle wpływa pozytywnie na poprawność funkcjonowania miar podobieństwa. Jednakże tego typu rozwiązania są znacznie trudniejsze i droższe w realizacji, gdyż wymagają zgromadzenia wiedzy z zakresu zgodnego z tematyką porównywanych tekstów oraz przyjęcia sposo-

bu jej reprezentacji. Z tego powodu podejście to nie ma również waloru uniwersalności. Do reprezentacji wiedzy dziedzinowej najczęściej stosuje się ontologie.

Zasadniczym celem niniejszej pracy jest próba oceny przydatności znanych z literatury miar podobieństwa tekstów. W kolejnych punktach artykułu najpierw przedstawione zostaną wybrane miary podobieństwa oparte ma macierzy częstości, wykorzystane do porównania dokumentów, a następnie dokumenty tekstowe, które były porównywane w badaniu. W dalszej części zaprezentowane będą wyniki przeprowadzonej analizy symulacyjnej opisanych wcześniej miar. Pracę zakończą wnioski z badań oraz krótki opis dalszych planów badawczych.

2. Miary podobieństwa tekstów oparte na macierzy częstości

W badaniu wzięto pod uwagę cztery miary podobieństwa stosowane w odniesieniu do tekstów. Dokonując wyboru miar podobieństwa do badania, uwzględniono wyniki badań przeprowadzonych przez dr. Dariusza Borratyńskiego (zob. jego rozprawę doktorską *Ocena przydatności częstotliwościowej reprezentacji dokumentów w języku polskim*). Wybrane miary podzielone zostały na dwie grupy: znormalizowane i nieznormalizowane.

Wśród znormalizowanych miar podobieństwa tekstów do badań symulacyjnych wybrano odległość cosinusową (wzór 1) oraz odległość Jaccarda (wzór 2). Przy wyznaczaniu odległości Jaccarda tworzone są dwa zbiory: X − będący zbiorem wyrazów występujących w pierwszym dokumencie oraz Y − zawierający wyrazy pochodzące z drugiego dokumentu. Pozwala to wyrazić odległość pomiędzy dokumentami za pomocą formuły odnoszącej część wspólną obu zbiorów do ich sumy w sposób przedstawiony w formule 2 [Deza, Deza 2009]:

$$d_1(X,Y) = 1 - \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2 \, y_k^2}}, \tag{1}$$

$$d_2(X,Y) = 1 - \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|}, \tag{2}$$

gdzie: X, Y − dokumenty, kolumny macierzy częstości,
$d(X,Y)$ − odległość pomiędzy dokumentami X i Y,
k − numery wyrazów, wiersze macierzy częstości,
x_k, y_k − liczba wystąpień *k*-tego słowa w dokumentach X, Y, elementy macierzy częstości na przecięciu *k*-tego wiersza i kolumn X i Y.

Dużą zaletą znormalizowanych miar jest łatwość przekształcenia miary odległości na miarę podobieństwa. Wykorzystany do tego może zostać wzór 3. Miary znormalizowane są również łatwiejsze w interpretacji. Można je wyrazić jako wartość procentowa, co bezpośrednio może zostać przełożone na stwierdzenie, że dokument X jest podobny do dokumentu Y w $100 \cdot s(X,Y)$ %.

$$s(X,Y) = 1 - d(X,Y), \tag{3}$$

gdzie $s(X,Y)$ – podobieństwo pomiędzy dokumentami X i Y.

Miary nieznormalizowane w badaniu reprezentuje odległość euklidesowa (wyrażona wzorem 4) oraz odległość miejska (przedstawiona wzorem 5).

$$d_3(X,Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}, \tag{4}$$

$$d_4(X,Y) = \sum_{k=1}^{n} |x_k - y_k|, \tag{5}$$

Miary te nie mogą zostać przekształcone na miary podobieństwa. Również ich interpretacja jest trudniejsza niż w przypadku miar znormalizowanych.

3. Analiza symulacyjna przydatności miar podobieństwa tekstów

Tę część artykułu poświecono omówieniu wyników badań symulacyjnych, dotyczących przydatności wybranych miar podobieństwa tekstów. Na początku przedstawione zostały dokumenty tekstowe, które były wykorzystywane w badaniu. W dalszej części znalazła się analiza wyników badań podzielona, podobnie jak w przypadku opisu miar, na wyniki dotyczące miar znormalizowanych i nieznormalizowanych. Jest to spowodowane głównie różnicą w interpretacji wyników.

3.1. Dokumenty tekstowe wykorzystane w badaniu

W symulacji wzięto pod uwagę pięć grup tekstów nazywanych kolekcjami. Poszczególne kolekcje zawierały po 20 dokumentów tekstowych o różnej długości i tematyce. W każdej grupie występował jeden tekst podstawowy i jeden tekst pomocniczy o tematyce różnej niż tekst podstawowy. Pozostałe dokumenty w grupie były różnymi modyfikacjami tekstu podstawowego, w niektórych przypadkach z domieszką tekstu pomocniczego. Strukturę tekstów we wszystkich grupach przedstawia druga kolumna w tabeli 1. W tabeli tej znajdują się również znane wartości podobieństwa pomiędzy dokumentami, wyznaczone na podstawie rodzaju (wycię-

cie fragmentu tekstu, wstawienie fragmentu tekstu podstawowego lub pomocniczego) i stopnia modyfikacji tekstu podstawowego.

Przed przystąpieniem do symulacyjnej analizy przydatności wybranych miar podobieństwa tekstów należy poddać dokumenty wstępnemu przetwarzaniu. W tym celu każda z badanych grup dokumentów została połączona w kolekcję dokumentów. Następnie na każdej kolekcji w ramach wstępnego przetwarzania dokumentów podjęte zostały takie działania, jak zamiana wszystkich liter na małe, usunięcie interpunkcji i białych znaków oraz usunięcie słów znajdujących się na tzw. stopliście.

Zarówno w badaniu symulacyjnym, jak i w przetwarzaniu wstępnym wykorzystano język R, a w szczególności pakiet tm [Feinerer 2010]. Obecnie w pakietach języka R, wspomagających eksploracyjną analizę tekstów, nie ma zdefiniowanej stoplisty dla języka polskiego. W literaturze również nie występuje standardowa stoplista dla języka polskiego. Z tego powodu na potrzeby badania została skonstruowana przykładowa stoplista. Ma ona częściowo subiektywny charakter i pozwala na usunięcie z badanych dokumentów wyrazów nieistotnych z punktu widzenia treści dokumentu. Na stopliście znalazły się takie wyrazy, jak: a, i, o, w, z, że, na, pod, obok, do, od, lub, ale, lecz, to, po, więc, czyli, który, nie, niż, kto, co, dla, jeszcze, jeśli, już, nad, by, ani. Wyrazy jedno- i dwuliterowe znalazły się tu tylko ze względów formalnych. Na dalszym etapie wyrazy te zostałyby automatycznie wyeliminowane.

W podobnych badaniach dla tekstów anglojęzycznych przeprowadza się dodatkowo redukcję wyrazów do rdzenia, czyli formy podstawowej, oraz zamienia określone terminy na ich synonimy. To ostatnie działanie wykonywane jest w języku R przy wykorzystaniu WordNetu [Feinerer i in. 2008].

Niestety, podobnie jak w przypadku stoplisty, żaden z pakietów języka R nie ma możliwości redukowania do rdzenia wyrazów tekstu w języku polskim. Podobnie jest w przypadku zamiany określonych terminów na synonimy. Tu dodatkowo dochodzi kwestia stopnia zaawansowania polskiego WordNetu, który mógłby tu zostać wykorzystany.

Wspomniane działania (redukcja do rdzenia i zamiana na synonimy) są niejednokrotnie bardzo istotne dla przeprowadzanej analizy. Przykładowo w systemach antyplagiatowych, gdzie eksploracyjna analiza tekstu może mieć zastosowanie, utrata informacji o kolejności wystąpienia wyrazów nie jest tak problematyczna jak pojawiające się w dokumentach synonimy. Najczęstsze praktyki plagiatowe w zakresie tekstów obejmują zmianę szyku zdania oraz stosowanie synonimów. Niemożność wychwycenia tego procederu może skutkować błędną oceną podobieństwa tekstów. Z tego powodu w dalszych badaniach konieczne będzie znalezienie innych narzędzi pozwalających przede wszystkim na redukcję wyrazów do rdzenia, a w dalszej kolejności na zamianę wybranych terminów na synonimy.

Na podstawie pięciu kolekcji dokumentów zostały dla każdej z nich utworzone dwie wersje macierzy częstości: podstawowa oraz binarna.

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1\,20} \\ a_{21} & a_{22} & \dots & a_{2\,20} \\ \vdots & \vdots & \ddots & \vdots \\ a_{w1} & a_{w2} & \dots & a_{w\,20} \end{bmatrix}_{w \times 20}, \tag{6}$$

gdzie: $A_k\,(k = 1, \dots, 5)$ − macierze częstości dla poszczególnych kolekcji dokumentów,

a_{ij} − liczba wystąpień *i*-tego ($i = 1, \dots, w$) wyrazu w *j*-tym $(j = 1, \dots, 20)$ dokumencie (w przypadku macierzy binarnej wystąpienie danego wyrazu w dokumencie ($a_{ij} = 1$) lub jego brak ($a_{ij} = 0$)),

w − liczba wyrazów w kolejnych kolekcjach dokumentów; dla kolejnych kolekcji wartości *w* wynoszą odpowiednio 1206, 1478, 583, 854 oraz 2523.

Macierz częstości to macierz, której kolumny reprezentują dokumenty, a wiersze − wyrazy (wzór 6). Wartości wewnątrz macierzy częstości w jej wersji podstawowej reprezentują liczbę wystąpień danego słowa w danym dokumencie, a w wersji binarnej wszystkie wartości niezerowe zostały zamienione na 1.

Na tak przygotowanych macierzach częstości można przeprowadzać dalsze badania.

3.2. Wyniki analizy symulacyjnej

Wyniki badania dotyczącego miar znormalizowanych przedstawia tab. 1. Wartości znajdujące się w niej to zamienione na podobieństwo wyrażone w procentach średnie odległości pomiędzy pierwszym dokumentem w każdej z kolekcji a pozostałymi dokumentami z danej kolekcji. Wartości zaznaczone ciemnoszarym kolorem to te, które najbardziej odpowiadają znanym wartościom podobieństwa pomiędzy tekstami. Osobno rozpatrywano tu wyniki dla różnych postaci macierzy częstości. Wartości oznaczone jasnoszarym kolorem były trudne do oceny ze względu ma znaczną domieszkę tekstu pomocniczego, nieznacznie podobnego do tekstu podstawowego.

Ogólnie można powiedzieć, że dla podstawowej postaci macierzy częstości miara cosinusowa sprawdza się lepiej, gdy powielany jest fragment tekstu głównego. Z kolei odległość Jaccarda daje lepsze rezultaty, gdy fragmenty tekstu są usuwane. Dla binarnej postaci macierzy częstości w większości przypadków lepiej lub co najmniej tak samo dobrze sprawdzała się miara Jaccarda. Jedynie przy zamianie fragmentów tekstu podstawowego na fragmenty tekstu pomocniczego wyniki bardziej zbliżone do znanego stopnia podobieństwa uzyskiwane były przy odległości cosinusowej.

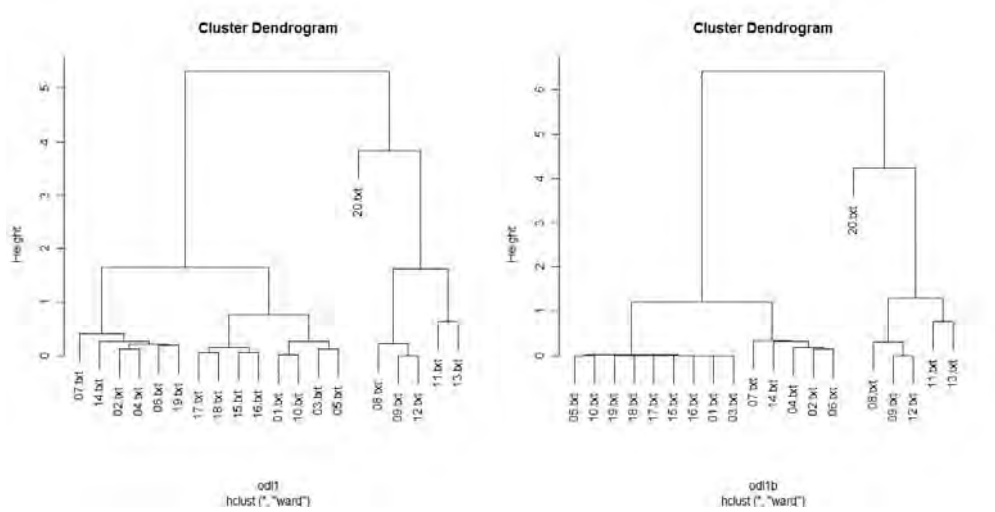
Tabela 1. Uśrednione wyniki badania symulacyjnego, zawierające rzeczywiste i wyznaczone miary podobieństwa pomiędzy dokumentem określanym jako tekst podstawowy a dokumentem powstałym w sposób opisany w kolumnie „dokument porównywany", wszystkich badanych kolekcji dokumentów dla miar znormalizowanych (w %)

Dokument porównywany	Znana wartość	Macierz częstości			
		podstawowa		binarna	
		cosinus	Jaccard	cosinus	Jaccard
Tekst podstawowy	100	100	100	100	100
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 10% całości	90	98	95	96	91
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 10% całości, został powtórzony	100	98	96	100	100
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 10% całości	90	98	95	96	92
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 10% całości, został powtórzony	100	98	96	100	100
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 20% całości	80	95	87	91	83
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 30% całości	70	92	80	86	74
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 40% całości	60	87	68	80	64
Tekst podstawowy po usunięciu losowo wybranego fragmentu o wielkości stanowiącej 50% całości	50	83	58	74	55
Podstawowy tekst	100	100	100	100	100
Połowa tekstu podstawowego z dodaną połową tekstu pomocniczego	50	67	50	55	37
Powtórzona dwukrotnie pierwsza połowa tekstu podstawowego	50	84	70	74	54
70% tekstu podstawowego z dodanymi 30% tekstu pomocniczego	70	82	69	72	56
90% tekstu podstawowego z dodanymi 10% tekstu pomocniczego	90	95	91	91	83
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 20% całości, został powtórzony	100	97	91	100	100
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 30% całości, został powtórzony	100	98	88	100	100
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 40% całości, został powtórzony	100	97	84	100	100
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 50% całości, został powtórzony	100	98	81	100	100
Tekst podstawowy, w którym losowo wybrany fragment, stanowiący 60% całości, został powtórzony	100	98	81	100	100
Pomocniczy tekst		20	11	9	5

Źródło: obliczenia własne.

Rysunek 1 przedstawia dwa przykładowe podziały dokumentów na klasy. Obydwa zostały dokonane na dokumentach tworzących pierwszy z analizowanych korpusów i dla odległości cosinusowej. Pierwszy podział bazuje jednak na podsta-

wowej macierzy częstości, a drugi na jej binarnej wersji. Na obydwu dendrogramach można zauważyć wyraźnie trzy grupy dokumentów oraz odstający od reszty dokument pomocniczy.



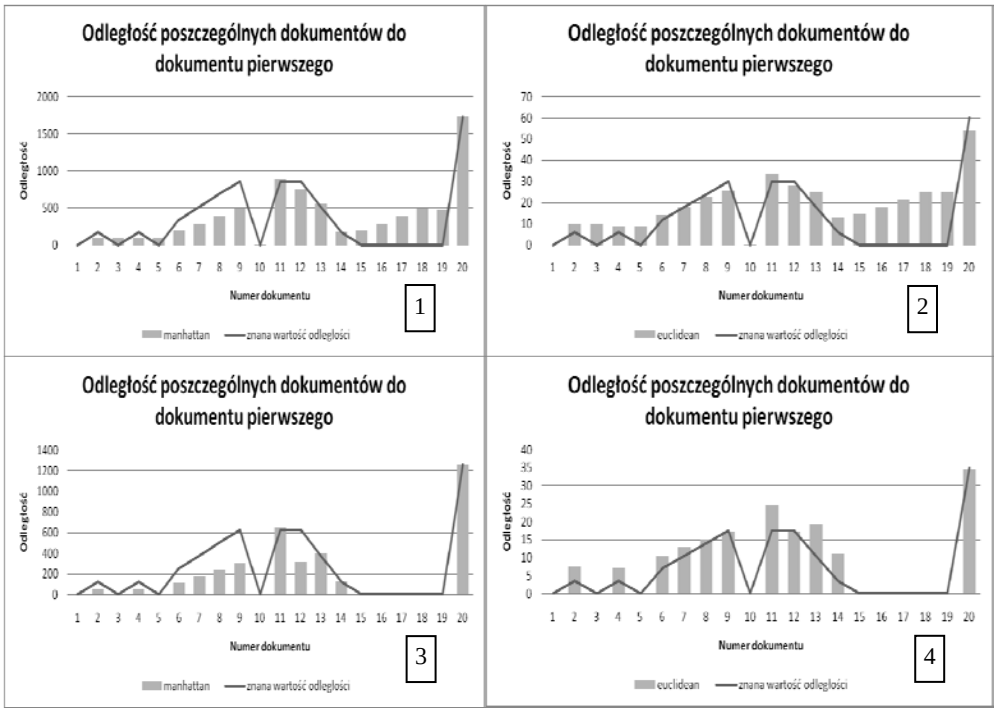
Rys. 1. Przykładowa klasyfikacja dokumentów metodą Warda przy wykorzystaniu cosinusowej miary odległości na macierzy częstości w wersji podstawowej (pierwszy dendrogram) i binarnej (drugi dendrogram) dla pierwszej kolekcji dokumentów

Źródło: opracowanie własne.

Kolejny etap analizy symulacyjnej to porównanie miar nieznormalizowanych. Wyniki analizy przedstawia rys. 2. Ponieważ wartości tych miar są uzależnione od rozmiaru dokumentów tekstowych, trudno jest oceniać wartości liczbowe. Trzeba również pamiętać, że w przeciwieństwie do miar znormalizowanych, tych nie da się przekształcić w miary podobieństwa. W tej sytuacji wyniki analizy przedstawione na wykresie pozwalają na dokładniejsze i bardziej pewne określenie przydatności każdej z tych miar. Podobnie jak w przypadku miar znormalizowanych, przedstawione wartości są średnią dla pięciu badanych kolekcji dokumentów.

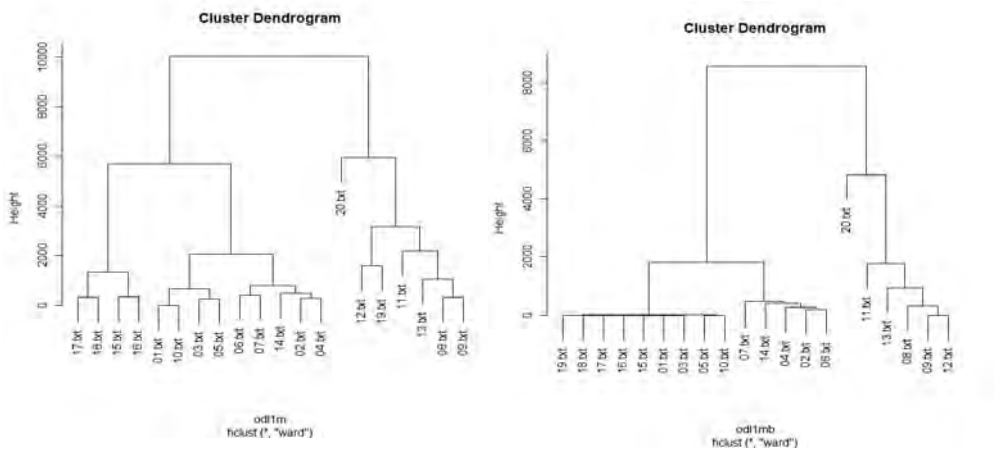
Z wykresów można wywnioskować, że miary bazujące na binarnej postaci macierzy częstości (wykresy 3 i 4 na rys. 2) dają lepsze wyniki. Z dwóch badanych miar odległość euklidesowa (wykresy 2 i 4 na rys. 2) wydaje się bardziej zgodna ze znanym poziomem odległości pomiędzy badanymi tekstami.

Podobnie jak w przypadku miar znormalizowanych, miary nieznormalizowane mogą stanowić podstawę podziału zbioru dokumentów na klasy. Przykładowe podziały dla odległości miejskiej przedstawia rys. 3, na którym również przedstawiono podział pierwszego zbioru dokumentów dla macierzy częstości w postaci podstawowej i binarnej.



Rys. 2. Uśrednione wyniki badania symulacyjnego dla miar znormalizowanych

Źródło: opracowanie własne.



Rys. 3. Przykładowy podział dokumentów na klasy metodą Warda przy wykorzystaniu odległości miejskiej na macierzy częstości w wersji podstawowej (pierwszy dendrogram) i binarnej (drugi dendrogram)

Źródło: opracowanie własne.

4. Podsumowanie

W artykule przedstawione zostały wyniki badań symulacyjnych dotyczących oceny przydatności wybranych miar podobieństwa tekstów. Trudno wybrać jedną miarę, której zastosowanie w każdej sytuacji byłoby najlepsze. W zależności od sytuacji wartość innej miary okazywała się najbliższa znanej wartości podobieństwa.

Dla podstawowej wersji macierzy częstości spośród miar znormalizowanych odległość cosinusowa daje wyniki bliższe do znanych wartości w przypadku powielania fragmentów tekstu. Odległość Jaccarda lepiej sprawdza się w sytuacjach, kiedy fragmenty tekstu zostają wycięte. Dla binarnej wersji macierzy częstości wyniki są podobne, ale z tą różnicą, że w przypadku powielania fragmentów tekstu obie miary dają takie same wyniki, a dodatkowo odległość cosinusowa sprawdza się lepiej w przypadkach dodawania do tekstu podstawowego fragmentów tekstu pomocniczego.

Spośród miar nieznormalizowanych badania wykazały większą skuteczność odległości euklidesowej. Ponadto wyniki zastosowania tej miary były bardziej zbliżone do znanych, gdy zastosowano binarną postać macierzy częstości.

W dalszych pracach planowane jest rozszerzenie zakresu stosowanych miar podobieństwa oraz zbadanie wpływu charakteru tekstu (ogólny, specjalistyczny) na uzyskiwane wyniki.

Literatura

Deza M.M., Deza E., *Encyclopedia of Distances*, Springer, Berlin − Heidelberg 2009.

Feinerer I., *Introduction to the tm Package. Text Mining in R*, http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf (10.12.2010).

Feinerer I., Hornik K., Meyer D., *Text mining infrastructure in R*, „Journal of Statistical Software", March 2008, vol. 25, issue 5.

Lula P., *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, http://www.statsoft.pl/czytelnia/8_2007/Lula05.pdf (10.12.2010).

COMPARATIVE ANALYSIS OF TEXT DOCUMENTS SIMILARITY MEASURES

Summary: The main objective of this paper is an attempt to evaluate the usefulness of known from literature similarity measures of text documents. In the subsequent parts of the paper there are chosen first measures based on frequency matrix which were used for the comparison of documents and next the text documents that were compared in the research. The next part of article is devoted to the results of previously presented methods' simulation analysis. On this basis the attempt to evaluate the usefulness of similarity measures of texts was made.