PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Paweł Lula
Akademia Ekonomiczna w Krakowie

KLASYFIKACJA DOKUMENTÓW TEKSTOWYCH SPORZĄDZONYCH W JĘZYKU POLSKIM

1. Wstęp

W zasobach informacyjnych gromadzonych i przetwarzanych przez człowieka przeważają dokumenty tekstowe. Upowszechnienie zasobów tekstowych i wynikająca z tego niemożność przeanalizowania zawartych w nich informacji stwarzają konieczność wypracowywania i udoskonalania metod, pozwalających na zautomatyzowanie ich przetworzenia. Realizacja tego zadania wymaga rozwiązania wielu problemów związanych z niskim stopniem strukturyzacji tekstów, wieloznacznością wypowiedzi i zróżnicowanymi cechami poszczególnych języków etnicznych.

Jedną z najważniejszych form analizy danych jest klasyfikacja. Metody klasyfikacji wykorzystywane były pierwotnie do analizy danych o charakterze numerycznym. Coraz mocniej zaznaczającej się potrzebie klasyfikacji danych o innym charakterze towarzyszyły próby dostosowania klasycznych metod taksonomicznych do realizacji ich nowych zadań. W większości przypadków jest stosowane podejście zakładające potrzebę przekształcenia zgromadzonych zasobów informacyjnych do postaci numerycznej i dalsze jej przetworzenie w sposób klasyczny. Idea ta jest również wykorzystywana w odniesieniu do klasyfikacji tekstów. Każdy dokument w pierwszym etapie analizy jest przekształcany do postaci wektora wartości numerycznych. Następnie zbiór wektorów, reprezentujący rozpatrywany zestaw dokumentów, jest analizowany w tradycyjny sposób.

Można wskazać na duże możliwości aplikacyjne taksonomicznej analizy dokumentów tekstowych. Należą do nich m.in. zagadnienia określania charakteru dokumentu oraz podobieństwa pomiędzy dokumentami.

Tematyka artykułu jest związana z zagadnieniami taksonomicznej klasyfikacji bezwzorcowej dokumentów tekstowych. Zasadniczym celem pracy jest prezentacja wyni-

ków badań wybranych aspektów klasyfikacji dokumentów sporządzonych w języku polskim. Artykuł ma następującą strukturę: omówienie procesu klasyfikacji bezwzorcowej tekstów wraz z przykładem zaprezentowano w punkcie drugim, charakterystyka przeprowadzonego eksperymentu badawczego i uzyskane wyniki stanowią istotę punktu trzeciego, pracę kończy podsumowanie wniosków wynikających z badań.

2. Klasyfikacja tekstów

Proces klasyfikacji dokumentów jest wieloetapowy. Do najważniejszych etapów należy zaliczyć:

- przygotowanie zbioru dokumentów,
- podział dokumentów na wyrazy,
- wstępne przetworzenie list wyrazów,
- wyznaczenie macierzy częstości (numeryczna reprezentacja dokumentów),
- przekształcenia macierzy częstości,
- zastosowanie metod taksonomicznych.

Zamieszczone w dalszej części artykułu punkty zawierają krótką charakterystykę poszczególnych etapów analizy.

2.1. Przygotowanie zbioru dokumentów

Podstawowym zadaniem pierwszego etapu analizy jest przeprowadzenie wstępnej obróbki analizowanego zbioru dokumentów. Obejmuje ona m.in.:

- transformację dokumentów do postaci tekstowej (znaczna liczba programów pozwalających na analizę tekstów operuje wyłącznie plikami tekstowymi),
- usunięcie znaków formatujących (np. usunięcie znaczników języka HTML z dokumentów pobranych z serwisów WWW),
- ujednolicenie sposobu kodowania znaków (etap ten jest bardzo istotny w przypadku tekstów polskich, gdyż sposób kodowania znaków charakterystycznych dla języka polskiego nie jest ujednolicony).

2.2. Podział dokumentów na wyrazy

Realizacja tego etapu analizy polega na usunięciu z każdego dokumentu znaków interpunkcyjnych i utworzeniu listy słów występujących w tekście.

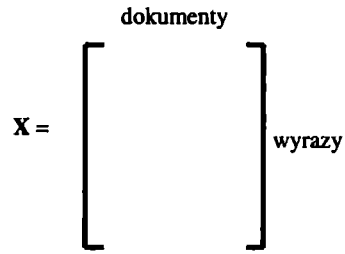
2.3. Wstępne przetworzenie list wyrazów

Listy wyrazów, związane z poszczególnymi dokumentami, są poddawane przetworzeniu, które zwykle obejmuje:

- usunięcie wyrazów nieistotnych z punktu widzenia dalszej analizy (wyrazy te tworzą tzw. *stop-listę*),
- przekształcenie wyrazów występujących na listach do ich formy podstawowej; transformacja do formy podstawowej jest określana również jako redukcja do rdzenia (ang. *stemming*); w zależności od specyfiki języka przekształcenie wyrazów do ich form podstawowych może być realizowane za pomocą reguł lub z wykorzystaniem słowników.

2.4. Wyznaczenie macierzy częstości

Wyrazy pochodzące z list odpowiadających poszczególnym dokumentom są łączone w jedną, wspólną listę. Następnie zlicza się, ile razy każdy wyraz wystąpił w każdym dokumencie. Zebrane dane tworzą *macierz częstości*, która w dalszej części tekstu oznaczana będzie symbolem **X** (rys. 1).



Rys. 1. Struktura macierzy częstości
Źródło: opracowanie własne.

Wierszom macierzy odpowiadają kolejne słowa występujące w dokumentach, kolumny zaś reprezentują dokumenty. Element macierzy x_{ij} określa liczbę wystąpień i-tego słowa w j-tym dokumencie. Analiza macierzy pozwala na badanie podobieństwa słów oraz dokumentów. Podobieństwo słów jest wyrażane poprzez określenie podobieństwa odpowiadających im wierszy macierzy **X**, o podobieństwie dokumentów wnioskuje się poprzez analizę podobieństwa kolumn tej samej macierzy.

Z punktu widzenia przeprowadzanej analizy szczególnie interesujące są te wyrazy, które występują w więcej niż jednym dokumencie, gdyż one wskazują na cechy wspólne rozpatrywanych tekstów. Z tego powodu w macierzy częstości uwzględnia się tylko te wiersze, które mają wartości niezerowe na więcej niż jednej pozycji. Wiersze odpowiadające wyrazom występującym tylko w jednym dokumencie są usuwane z macierzy.

2.5. Przekształcenia macierzy częstości

Wyznaczone częstości wystąpień poddać można transformacji, służącej uwypukleniu cech wspólnych dokumentów. Potrzebę przeprowadzenia transformacji uzasadniać może spostrzeżenie mówiące, że o stopniu podobieństwa pomiędzy doku-

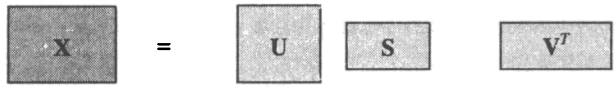
mentami mocniej świadczy samo wystąpienie takich samych słów od precyzyjnie określonej liczby ich wystąpień.

Do najczęściej stosowanych metod transformacji macierzy częstości należą:

1. Reprezentacja binarna – zakłada, że rejestrowany jest sam fakt wystąpienia i-tego słowa w j-tym dokumencie, nie precyzuje się zaś liczby wystąpień. Zastosowanie tej metody prowadzi do zastąpienia oryginalnej macierzy częstości macierzą wystąpień o elementach x_{ij} równych:
   - jedności, jeśli i-te słowo występuje w j-tym dokumencie (jeden bądź więcej razy),
   - zero, jeśli i-te słowo nie występuje w j-tym dokumencie.
2. Reprezentacja logarytmiczna – polega na zastąpieniu wszystkich niezerowych elementów macierzy **X** wartościami równymi $\left(1+\log\left(x_{ij}\right)\right)$. Elementy zerowe macierzy częstości nie ulegają zmianie. Stosowanie reprezentacji logarytmicznej ma podobne uzasadnienie jak użycie reprezentacji binarnej: uwypukla ona wystąpienie słowa w dokumencie i tylko w niewielkim stopniu uwzględnia liczbę wystąpień tego wyrazu.
3. Ważona reprezentacja logarytmiczna – podstawą jej zdefiniowania było spostrzeżenie, że jeśli pewien wyraz występuje w każdym rozpatrywanym dokumencie, to nie pozwala on na rozróżnienie i grupowanie tekstów, z punktu widzenia możliwości klasyfikacji szczególnie przydatne są zaś wyrazy występujące w stosunkowo niewielkiej liczbie dokumentów. Podejście takie znajduje swoje odzwierciedlenie w stosowaniu formuły [Manning i in. 1999]: $\left(\left(1+\log\left(x_{ij}\right)\right)*\log\left(\frac{N}{df_i}\right)\right)$ dla elementów niezerowych macierzy częstości i pozostawienie elementów zerowych na dotychczasowym poziomie (w przedstawionej formule wartość N określa liczbę wszystkich dokumentów, df_i jest zaś liczbą dokumentów zawierających i-ty wyraz.

Jedną z trudności, pojawiających się w dalszej analizie macierzy częstości (w postaci pierwotnej lub przekształconej), jest jej duży rozmiar. Szczególnie uciążliwa jest liczba wierszy równa liczbie wszystkich różnych wyrazów występujących w badanych dokumentach. Powszechnie stosowanym środkiem zaradczym jest zastosowanie metody LSA (*latent semantic analysis*) zaproponowanej w pracy [Deerwester i in. 1990]. Opiera się ona na zastosowaniu w odniesieniu do macierzy częstości (w postaci oryginalnej lub przekształconej) rozkładu według wartości osobliwych (schemat pokazuje rys. 2).

Celem obliczeń jest zdefiniowanie przestrzeni, w której byłaby możliwa analiza zbiorów wyrazów występujących w dokumentach (wyrazy reprezentuje macierz **U**) oraz analiza zbioru dokumentów (macierz **V**). Każdy obiekt (wyraz, dokument) jest reprezentowany przez jeden wiersz odpowiedniej macierzy, przy czym (podobnie jak w analizie głównych składowych) poszczególne współrzędne są uporządkowane malejąco według ich wartości informacyjnej, co pozwala na redukcję przestrzeni poprzez uwzględnienie tylko pewnej liczby początkowych elementów wektora.



macierz U – wyrazy w przestrzeni wyznaczonej przez składowe

macierz V – dokumenty w przestrzeni wyznaczonej przez składowe

macierz S – macierz diagonalna, znaczenie kolejnych składowych

Rys. 2. Schemat rozkładu macierzy częstości według wartości osobliwych
Źródło: opracowanie własne.

W klasyfikacji dokumentów przedstawiony schemat postępowania zastosowano w odniesieniu do macierzy **V**, której elementy, po przeskalowaniu przez współczynniki zawarte w macierzy **S**, wyznaczają punkty reprezentujące dokumenty uwzględnione w badaniu. Z praktycznego punktu widzenia szczególnie przydatna byłaby redukcja przestrzeni, w której opisywane są dokumenty, do dwóch lub trzech wymiarów. Pozwoliłoby to na graficzną reprezentację struktury zbioru dokumentów.

2.6. Zastosowanie metod taksonomicznych

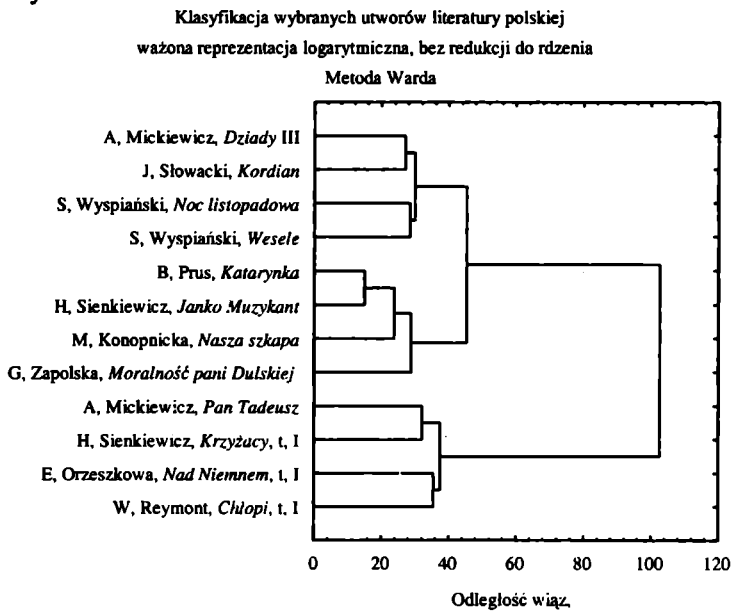
Po wyznaczeniu ostatecznej postaci macierzy, stanowiącej numeryczną reprezentację badanego zbioru dokumentów, możliwe jest zastosowanie metod taksonomicznych do przeprowadzenia stosowanych obliczeń potrzebnych do dokonania klasyfikacji.

2.7. Przykład klasyfikacji polskich tekstów

W celu zilustrowania przedstawionej procedury klasyfikacji dokonano analizy taksonomicznej tekstów wybranych dzieł literatury polskiej. Analiza obejmowała następujące pozycje:

- Adam Mickiewicz, *Dziady III*,
- Juliusz Słowacki, *Kordian*,
- Stanisław Wyspiański, *Noc listopadowa*,
- Stanisław Wyspiański, *Wesele*,
- Bolesław Prus, *Katarynka*,
- Henryk Sienkiewicz, *Janko Muzykant*,
- Maria Konopnicka, *Nasza szkapa*,
- Gabriela Zapolska, *Moralność pani Dulskiej*,
- Adam Mickiewicz, *Pan Tadeusz*,
- Henryk Sienkiewicz, *Krzyżacy* (t. I),
- Eliza Orzeszkowa, *Nad Niemnem* (t. I),
- Władysław Reymont, *Chłopi* (t. I).

Zastosowanie metody Warda pozwoliło na uzyskanie dendrogramu zaprezentowanego na rys. 3.



Rys. 3. Wyniki klasyfikacji treści wybranych polskich utworów literackich
Źródło: opracowanie własne.

W trakcie obliczeń wykorzystano oprogramowanie napisane w języku STATISTICA Visual Basic oraz pakiet STATISTICA wraz z modułem Text Miner. Uzyskany dendrogram pokazuje podobieństwo pomiędzy treścią analizowanych utworów.

3. Charakterystyka eksperymentu badawczego

3.1. Cele badań

W badaniach przyjęto następujące cele i założenia:

- badania dotyczyły klasyfikacji dokumentów tekstowych sporządzonych w języku polskim,
- badano przydatność dwóch metod klasyfikacji: metody *k*-średnich i algorytmu Warda,
- rozważono przydatność czterech sposobów reprezentacji dokumentów: macierzy częstości, reprezentacji binarnej, logarytmicznej i logarytmicznej ważonej,
- rozważano przydatność rozkładu według wartości osobliwych,
- określono przydatność stosowania stop-listy,
- badano wpływ redukcji wyrazów do formy podstawowej (redukcji do rdzenia).

3.2. Charakterystyka danych i przebieg eksperymentu

Do eksperymentu wykorzystano 10 artykułów pobranych z serwisu www.onet.pl. Dotyczyły one czterech zagadnień:

- obecnej sytuacji w Iraku (3 artykuły),
- powodzi w Kornwalii (2 artykuły),
- ściągalności opłat za abonament RTV (2 artykuły),
- doniesień z odbywającej się w trakcie badań olimpiady w Atenach (3 artykuły).

W obliczeniach[1] wykonano wiele eksperymentów symulacyjnych. W trakcie każdego z nich wartości jednego z rozważanych parametrów ustalano na stałym poziomie, a wartości wszystkich pozostałych parametrów modyfikowano w dopuszczalnym dla nich zakresie. W ten sposób określano wpływ rozważanego parametru na wyniki klasyfikacji. Do oceny poprawności klasyfikacji zastosowano odsetek prawidłowych klasyfikacji dokonanych przy ustalonej wartości rozważanego parametru.

3.3. Uzyskane wyniki

Do najciekawszych rezultatów uzyskanych w trakcie obliczeń należy zaliczyć rozważania określające:

- wpływ postaci macierzy częstości

Metoda	Prawidłowa klasyfikacja (%)
Macierz częstości	0,00
Reprezentacja binarna	58,33
Reprezentacja logarytmiczna	37,50
Ważona reprezentacja logarytmiczna	54,17

- wpływ stop-listy

Zastosowanie stop-listy	Prawidłowa klasyfikacja (%)
Nie	25,00
Tak	43,75

- wpływ zastosowania redukcji do rdzenia

Redukcja do rdzenia	Prawidłowa klasyfikacja (%)
Nie	35,94
Tak	40,63

- wpływ zastosowania metody LSA

Zastosowanie LSA	Prawidłowa klasyfikacja (%)
Nie	58,33
Tak, wszystkie 10 składowych	33,33
Tak, tylko 2 składowe	0,00
Tak, tylko 5 składowych	58,33

[1] Wszystkie obliczenia zrealizowano za pomocą programów pisanych w języku Perl oraz Visual Basic. Wykorzystano również pakiet STATISTICA wraz z modułem Text Miner oraz analizator morfologiczny SAM.

- wpływ wybranej metody taksonomicznej

Metoda	Prawidłowa klasyfikacja (%)
k-średnich	31,25
Warda, kwadrat odległość Euklidesa	43,75

4. Wnioski

Wydaje się, że przeprowadzony eksperyment pozwala na sformułowanie następujących wniosków:

- wskazane jest stosowanie ważonej reprezentacji logarytmicznej lub binarnej,
- wskazane jest stosowanie stop-listy,
- metoda Warda dawała lepsze wyniki niż metoda *k*-średnich,
- przeprowadzenie redukcji do rdzenia nie wpływa w istotny sposób na poprawę wyników,
- wskazane jest stosowanie rozkładu SVD, ale przy zastosowaniu ważonej reprezentacji logarytmicznej lub binarnej,
- uwzględnienie dwóch składowych SVD nie prowadzi do uzyskania poprawnych wyników (0% poprawnych klasyfikacji),
- uwzględnienie 50% składowych prowadzi do lepszych rezultatów niż zastosowanie ich wszystkich; w zastosowaniu układu: ważona reprezentacja logarytmiczna, stop-lista, 50% składowych wszystkie uzyskane klasyfikacje były prawidłowe.

Sformułowane wnioski z przeprowadzonego eksperymentu znajdują również swoje potwierdzenie w innych badaniach przeprowadzonych przez autora. W wielu przypadkach podobne rezultaty uzyskano również dla innych języków europejskich. Należy podkreślić, że z powodu ograniczonego zakresu przeprowadzonych obliczeń zagadnienia poruszone w pracy wymagają dalszych badań.

Literatura

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R., *Indexing by Latent Semantic Analysis*, „Journal of the American Society for Information Science" 1990, 41 (6), s. 391-407.

Lovins J.B., *Development of a Stemming Algorithm*, „Mechanical Translation and Computational Linguistics" 1968, 11.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge 1999.

THE CLUSTER ANALYSIS OF POLISH TEXT DOCUMENTS

Summary

The main purpose of the paper is to study the process of cluster analysis of Polish text documents. We consider the influence of different factors (choice of methods of text representation, utilization of stop-list and stemming, conducting Latent Semantic Analysis and choice of clustering methods) on results of clustering procedure. At the final part of the paper on the basis of experiment results we formulate some recommendations for carrying out clustering analysis of Polish texts.