PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 — 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu

TECHNIKI WIZUALIZACJI TEKSTU Z WYKORZYSTANIEM CHMUR ZNACZNIKÓW I DRZEW ZNACZNIKÓW*

1. Wstęp

Tradycyjne techniki wizualizacji w szeroko rozumianej wielowymiarowej analizie statystycznej jako dane wejściowe przyjmują dane liczbowe w postaci macierzy danych.

Pewnym odstępstwem od tej reguły była technika *zoom-star* [Bock, Diday 2000] umożliwiająca prezentację również danych symbolicznych opisanych zmiennymi w postaci przedziałów liczbowych, listy kategorii oraz listy kategorii z wagami. Narzędzie to nie pozwalało jednak na wizualizację „czystych" danych testowych, co umożliwiają powstałe w ostatnich kilku latach i gwałtownie zyskujące na popularności takie techniki wizualizacji, jak chmury znaczników (*tag clouds*), chmury słów (*word clouds*) i drzewa znaczników (*tree clouds*).

Artykuł składa się z czterech części: pierwsza opisuje techniki wizualizacji za pomocą chmur znaczników, druga – chmur słów wraz z przykładami ich konkretnych realizacji, trzecia opisuje stosunkowo nową ich odmianę – drzewa znaczników, a czwarta proponuje obszary zastosowań tych technik w badaniach marketingowych i przedstawia uwagi końcowe.

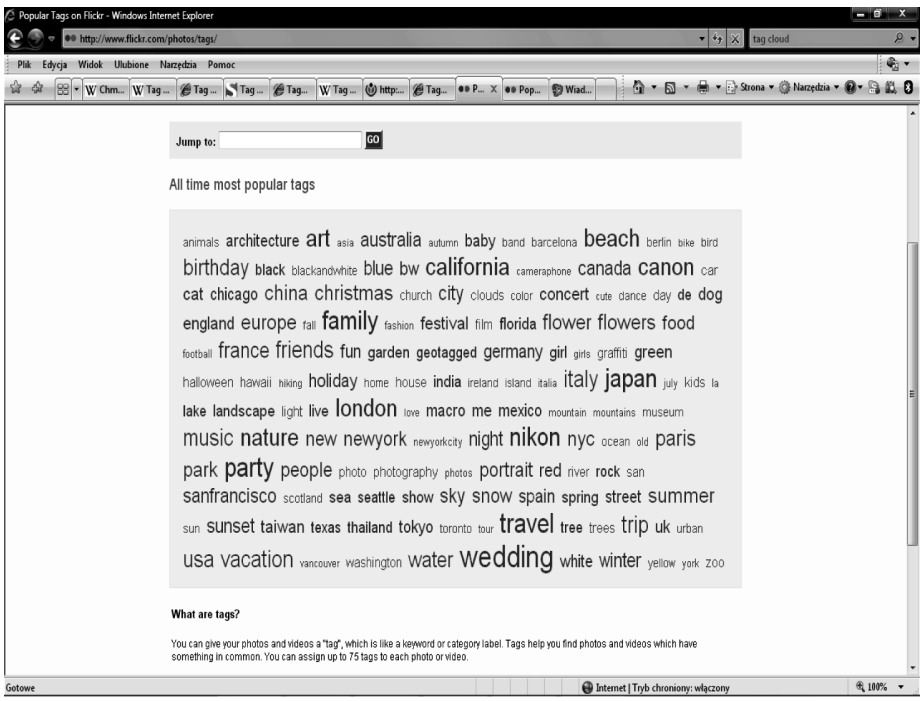
2. Chmury znaczników

Chmura znaczników (za http://pl.wikipedia.org/wiki/Chmura_znaczników) to graficzne zobrazowanie zawartości np. serwisu internetowego w postaci zestawu znaczników. Najczęściej znaczniki są uszeregowane alfabetycznie, natomiast ich wielkość i pogrubienie czcionki jest zależne od ważności lub popularności danego znacznika. Umożliwia to łatwe znalezienie danej kategorii zarówno alfabetycznie, jak i według ważności.

* Artykuł powstał w ramach projektu badawczego MNiSW N N111 105234 „Obiekty symboliczne w wielowymiarowej analizie statystycznej".

Chmury znaczników to narzędzie ściśle związane z rozwojem sieci Internet. Ta technika wizualizacji odzwierciedla, jak dużo obiektów (tekstu, grafiki, strumieni audio lub wideo) znajduje się na stronie lub jak popularne (jak często odwiedzane przez użytkowników) są poszczególne obiekty na stronie. Obiekty są powiązane z jednym znacznikiem lub kilkoma znacznikami za pomocą instrukcji języka HTML: `<Meta keywords="znacznik1;znacznik2;...">` lub, w przypadku autorskich systemów typu CMS (*Content Menagement Systems* – systemów zarządzania treścią stron WWW), za pomocą pól powiązanych z obiektem w bazie danych związanej z systemem. Chmura znaczników reprezentuje znaczniki przypisane do wszystkich obiektów, a jeśli strona zawiera dużo obiektów opisanych różnymi znacznikami, te znaczniki, które łącznie występują w liczbie większej niż pewna określona wartość progowa.

Jako pierwszy chmurę znaczników zastosował portal www.flickr.com oferujący umieszczanie zdjęć i pozwalający na opisywanie każdego z opublikowanych zdjęć słowami kluczowymi. Chmura znaczników reprezentuje najpopularniejsze słowa kluczowe używane do opisu wszystkich zdjęć. Chmura znaczników w tym wypadku jest interaktywna, a kliknięcie na któryś z nich powoduje wyświetlenie listy wszystkich zdjęć, w których opisie znajduje się to słowo kluczowe.



Rys. 1. Chmura znaczników na stronie www.flickr.com uważanej za pierwszy przykład zastosowania tej techniki w sieci Internet

Źródło: zrzut ekranu strony internetowej http://www.flickr.com/.

Algorytm tworzenia chmur znaczników jest stosunkowo prosty. W pierwszym kroku każdemu znacznikowi przypisuje się współczynnik w_z oznaczający liczbę jego wystąpień na stronie i wszystkich podstronach danej witryny WWW. W kroku drugim wszystkie znaczniki lub znaczniki, dla których w_z jest większy niż określona wartość progowa w_z (np. takie znaczniki, że obiekty z nimi związane zostały odwiedzone przez użytkowników co najmniej 100 razy), są wypisywane, zazwyczaj w kolejności alfabetycznej, przy czym wielkość czcionki jest proporcjonalna do w_z. W różnych odmianach tej techniki, zamiast zróżnicowania rozmiaru czcionki, wskaźnik w_z może mieć wpływ na intensywność wytłuszczenia czcionek, jej krój czy kolor. Rysunek 2 przedstawia różne odmiany chmur znaczników. Przykłady te pochodzą z listy „The TagCloud Top 100 List" z witryny www.tagcloud.com.

Rys. 2. Różne odmiany chmur znaczników

Źródło: opracowanie własne na podstawie „The TagCloud Top 100 List" (http://www.tagcloud.com/).

3. Chmury słów

Chmury słów (*word clouds*) są narzędziem pokrewnym chmur znaczników, jednak niezwiązanym tak ściśle z witrynami internetowymi. W przeciwieństwie do chmur znaczników danymi wejściowymi nie są znaczniki i powiązane z nimi obiekty internetowe, ale słowa ze „statycznych" dokumentów. Chmury słów nie mogą więc z natury rzeczy reprezentować popularności poszczególnych obiektów, ale najczęściej przedstawiają, ile razy dane słowo występuje w dokumencie lub zestawie dokumentów. Zdarzają się jednak wyjątki od tej reguły, jak np. chmura słów przedstawiająca liczbę ludności w największych krajach świata (rys. 3).

Rys. 3. Chmura słów przedstawiająca liczbę ludności największych krajów świata

Źródło: http://commons.wikimedia.org/wiki/File:World_Population.png.

Rys. 4. Chmura słów przedstawiająca słowa najczęściej występujące w streszczeniach referatów z XIII Warsztatów Metodologicznych (II Seminarium im. Profesora Stefana Mynarskiego) pt. „Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania"

Źródło: opracowanie własne, za pomocą aplikacji *Wordle* (http://www.wordle.com/).

Inną różnicą pomiędzy chmurami znaczników a chmurami słów jest to, że te pierwsze są zazwyczaj interaktywne, z możliwością klikania na poszczególne znaczniki, czego są pozbawione te drugie. Z tego też powodu w przypadku chmur słów wyrazy są znacznie rzadziej uporządkowane w kolejności alfabetycznej, a często ta kolejność jest dobierana losowo lub tak, aby uatrakcyjnić samą wizualizację. Przykładem takiego zastosowania tej techniki jest utworzony za pomocą popularnego narzędzia *Wordle* (http://www.wordle.com/) wykres przedstawiający słowa najczęściej występujące w streszczeniach referatów z XIII Warsztatów Metodologicznych (II Seminarium im. Profesora Stefana Mynarskiego) pt. „Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania", zaprezentowany na rys. 4.

4. Drzewa znaczników

Rozwinięciem pomysłu chmur znaczników i chmur słów są drzewa znaczników (*tree clouds*). W przypadku chmur znaczników wzajemne położenie słów nie odzwierciedla związku między nimi, a wynika z kolejności alfabetycznej (lub losowej). Drzewa znaczników, oprócz wizualizacji częstotliwości występowania poszczególnych słów, przedstawiają również związki pomiędzy nimi. Odległość pomiędzy poszczególnymi słowami na rysunku odzwierciedla to, jak często słowa te wchodzą w związki frazeologiczne lub występują w tych samych zdaniach.

Algorytm tworzenia drzew znaczników składa się z trzech etapów:

1. Obliczenie odległości pomiędzy słowami. Może do tego celu służyć jedna z odległości (według Gambette'a i Veronisa):

a) *mutual information distance*,

b) odległość Liddela,

c) odległość Jaccarda,

d) odległość hyperlex,

e) średnia geometryczna,

f) logarytm wiarygodności (*log likelihood*),

g) miara Poissona-Stirlinga.

2. Normalizacja wartości otrzymanych w kroku 1.

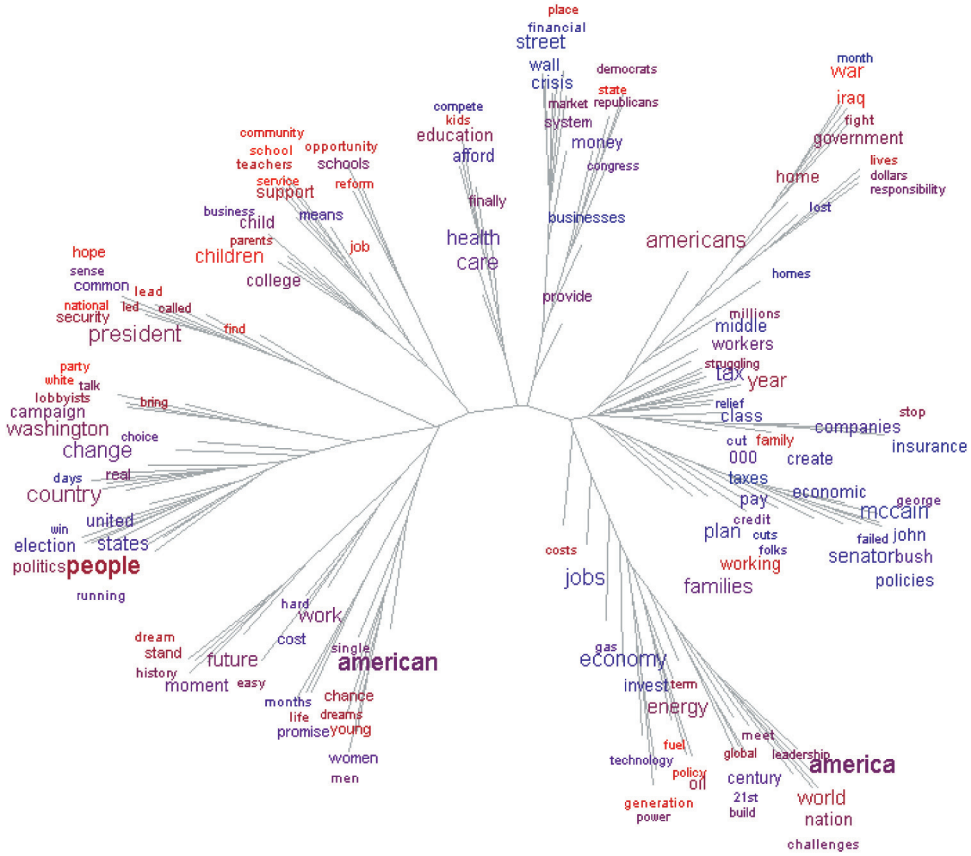
3. Utworzenie drzewa jedną z metod:

a) łączenia sąsiadów (*neighbor-joining*) [Saitou, Nei 1987],

b) wariantów dodawania do drzewa (*addtree variants*) [Barthelemy, Luong 1987],

c) heurystyki kwartetów (*quartet heuristic*) [Cilibrasi, Vitanyi 2007].

Rysunek 5 przedstawia drzewo znaczników powstałe poprzez analizę wystąpień Baracka Obamy w kampanii prezydenckiej przed wyborami w 2008 r., z wykorzystaniem odległości Jaccarda oraz metody łączenia sąsiadów do właściwego utworzenia drzewa.

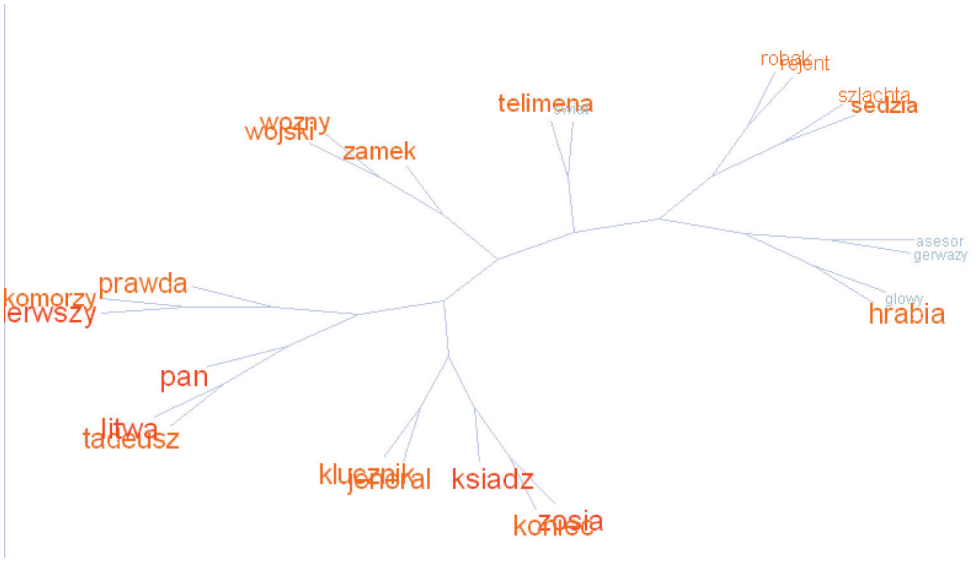


Rys. 5. Drzewo znaczników powstałe na podstawie przemówień z kampanii wyborczej Baracka Obamy w latach 2007-2008

Źródło: [Gambette, Veronis].

Dodatkowo na rys. 5 nasycenie koloru jest powiązane z czasem, z którego pochodzą przemówienia. Wyrazy w odcieniach czerwieni pochodzą z wcześniejszych przemówień, a wyrazy w odcieniach niebieskiego to te, które były najczęściej używane pod koniec kampanii.

Drzewa znaczników jako stosunkowo nowa technika wizualizacji nie doczekały się jeszcze implementacji w komercyjnych programach statystycznych ani w popularnym środowisku **R**. W celu ich utworzenia można jednak skorzystać z programu TreeCloudEng.exe, rozprowadzanego na zasadach licencji *Open Source*, dostępnego pod adresem http://www.lirmm.fr/~gambette/ProgTreeCloudENG.php. Na rysunku 6 znajduje się drzewo znaczników wygenerowane na podstawie zawartości „Pana Tadeusza" z wykorzystaniem odległości Jaccarda oraz metody łączenia sąsiadów.



Rys. 6. Drzewo znaczników powstałe na podstawie zawartości „Pana Tadeusza”

Źródło: opracowanie własne na podstawie [Mickiewicz 1834].

5. Obszary zastosowań i uwagi końcowe

Prezentowane w niniejszym artykule metody wizualizacji danych tekstowych są stosunkowo nową techniką i ich zastosowanie w szeroko rozumianym marketingu, a zwłaszcza w badaniach marketingowych, nie ma silnej podbudowy teoretycznej. Wydaje się jednak, że oprócz niewątpliwych zastosowań komercyjno-reklamowych można zaproponować kilka obszarów związanych z badaniami marketingowymi, w których omawiane techniki mogłyby z powodzeniem być wykorzystywane w procedurze badawczej:

- Wizualizacja badań ankietowych z pytaniami opisowymi. Pytania opisowe to cenne źródło danych, jednak ich kodowanie do słabych skal pomiarowych i ewentualne dalsze przetwarzanie jest zawsze związane z pewną utratą informacji. Wydaje się więc, iż zbiorcza wizualizacja tych opisów może być bardziej wartościowa dla badacza, a zwłaszcza dla odbiorcy badań, niż „suche” liczby otrzymane w tradycyjnych metodach wielowymiarowej analizy statystycznej.
- Opisy profili rynkowych.
- Mapy percepcji (np. w ocenie skuteczności kampanii marketingowych). Połączenie przez jedną z tych technik opinii potencjalnych klientów o produkcie po zakończonej kampanii mogłoby być bardzo użytecznym narzędziem w stwierdzeniu, czy przyniosła ona odpowiedni efekt i spowodowała, że produkt stał się rozpoznawalny i zajął zamierzone miejsce w świadomości przyszłych jego nabywców.

Literatura

Barthelemy J.P., Luong N.X., *Sur la topologie d'un arbre phylogenetique: aspects theoriques, algorithmes et applications l'analyse de donnees textuelles*, „Mathematiques et Sciences Humaines" 1987, 100, p. 57-80.

Bock H.-H., Diday E. (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg 2000.

Cilibrasi R., Vitanyi P. , *The Google Similarity Distance*, IEEE/ACM Transactions on Knowledge and Data Engineering, 2007.

Fujimura K., Fujimura S., Matsubayashi T., Yamada T., Okuda H., *Topigraphy: Visualization for Large-scale Tag Clouds*, Proceedings of the 17th International Conference on World Wide Web, 2008.

Gambette P., Veronis J., *Visualising a Text with a Tree Clouds*, Proceedings of XI International Federation of Classification Conference, w druku.

Gascuel O., Levy D., *A Reduction Algorithm for Approximating a (Nonmetric) Dissimilarity by a Tree Distance*. Journal of Classification 1996, 13(1), p. 129-155.

Hassan-Montero Y., Herrero-Solana V., *Improving Tag-Clouds as Visual Information Retrieval Interfaces*, InSciT2006.

Kaser O., Lemire D., *TagCloud Drawing: Algorithms for Cloud Visualization*, Proceedings of the 16th International Conference on World Wide Web, 2007.

Mickiewicz A., *Pan Tadeusz, czyli Ostatni Zajazd na Litwie. Historja szlachecka z r. 1811 i 1812, we dwunastu księgach wierszem*, Paryż 1834.

Saitou N., Nei M., *The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees*, Molecular Biology and Evolution 1987, 4, p. 406-425,

Veronis J., *Hyperlex, Lexical Cartography for Information Retrieval*, Computer, Speech and Language 2004, 18(3), p. 223-252.

TEXT VISUALIZATION TECHNIQUES WITH THE USE OF TAG CLOUDS AND TREE CLOUDS

Summary

Traditional data visualization techniques in a widely-mentioned multivariate statistical analysis require input in form of numerical data, most often – data matrices. The visualization of text data has never been considered in literature of subject and wording "text visualization" has been treated as an oxymoron.

The situation has changed due to the expansion of the Internet and growing in exponential rate, number of world-wide-web pages. In the last few years such techniques as tag clouds, word clouds and tree clouds have appeared and gained a rapid popularity.

In the article, creation of tag clouds, tag trees and tree clouds algorithms are described along with the various forms of tag clouds realization examples. The main areas of the usage of those techniques in marketing research are also proposed.