PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 328

Taksonomia 23
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 328 • 2014
Taksonomia 23 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Katarzyna Wójcik, Janusz Tuchowski
Uniwersytet Ekonomiczny w Krakowie

DOBÓR OPTYMALNEGO ZESTAWU SŁÓW ISTOTNYCH W OPINIACH KONSUMENTÓW NA POTRZEBY ICH AUTOMATYCZNEJ ANALIZY

Streszczenie: Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na rozwój działalności biznesowej. Wiele podejść do automatycznej analizy opinii konsumenckich opartych jest na podobieństwie lub też odległości między dokumentami tekstowymi. Powstaje więc problem, która z miar podobieństwa dokumentów tekstowych (lub ich odległości) jest najlepsza w przypadku charakterystycznego rodzaju tekstów, jakimi są opinie konsumentów. Głównym celem pracy jest przeprowadzenie analizy doboru optymalnego zestawu słów istotnych podczas obliczania podobieństwa (odległości) dokumentów tekstowych stosowanych na potrzeby automatycznej analizy opinii konsumenckich. Do obliczeń wykorzystane zostały: język R, program RapidMiner oraz arkusz kalkulacyjny.

Słowa kluczowe: *text mining*, *Web mining*, taksonomia, klasyfikacja dokumentów tekstowych, ocena jakości miar.

1. Wstęp

Analiza opinii konsumenckich jest obszarem badań, który może mieć znaczący wpływ na rozwój działalności biznesowej. Znaczna liczba konsumentów przed dokonaniem wyboru towaru lub usługi przeszukuje Internet, zaznajamiając się z opiniami innych użytkowników sieci. Znalezione rekomendacje często odgrywają decydującą rolę podczas podejmowania decyzji. Z kolei po dokonaniu zakupu lub skorzystaniu z usługi użytkownik jest proszony o wystawienie opinii lub też dokonuje tego z własnej inicjatywy. Narastająca liczba opinii dostępnych w sieci wytworzyła potrzebę ich automatycznej analizy i przetwarzania.

Wiele podejść do automatycznej analizy opinii konsumenckich opiera się na analizie podobieństwa lub też odległości pomiędzy dokumentami tekstowymi. Powstaje zatem problem, która z miar podobieństwa dokumentów tekstowych (lub ich odległości) jest najlepsza w przypadku charakterystycznego rodzaju tekstów, jakimi są opinie konsumentów.

Głównym celem pracy jest przeprowadzenie analizy doboru optymalnego zestawu słów istotnych podczas obliczania podobieństwa (lub odległości) dokumentów tekstowych, stosowanego na potrzeby automatycznej analizy opinii konsumenckich. Działanie takie pozwoli na zwiększenie efektywności badań poprzez zawężenie analizowanych danych. Wyeliminowane z analizy zostaną słowa nieistotne, niemające wpływu na wynik końcowy.

2. Automatyczna analiza opinii konsumenckich i miary podobieństwa tekstów

Automatyczna analiza opinii konsumenckich (*sentiment analysis*, *opinion mining*) to ogół działań mających na celu zautomatyzowanie procesu wyszukiwania, ekstrakcji i analizy danych pochodzących ze specyficznych tekstów, jakimi są opinie użytkowników. Są to działania z pogranicza przetwarzania języka naturalnego (*Natural Language Processing* – NLP), lingwistyki komputerowej (*computational linguistics*) oraz eksploracyjnej analizy tekstu (*text mining*). Jej celem jest określenie nastawienia autora wypowiedzi do jej przedmiotu [Wikipedia].

W ramach automatycznej analizy opinii konsumenckich wyróżnić można trzy rodzaje działań [Liu 2007; Pang, Lee 2008]:

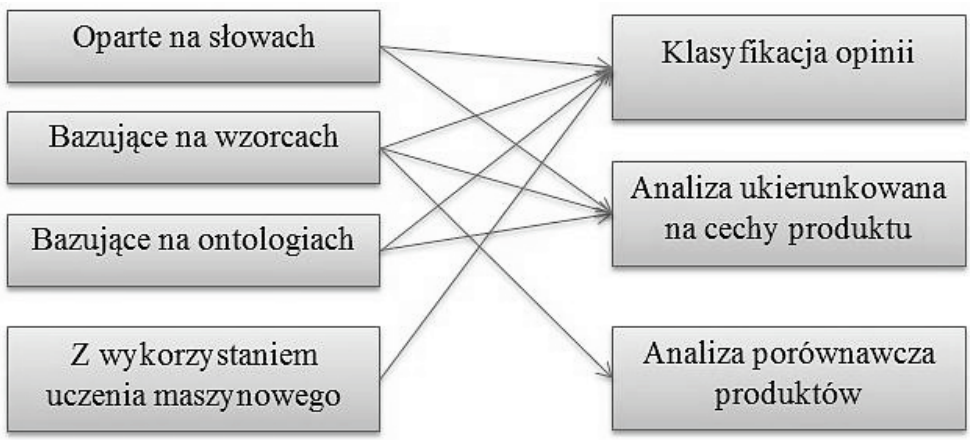
- **Klasyfikacja opinii** – podział opinii na grupy według ich nacechowania (np. pozytywne, negatywne, neutralne) lub przypisanie pojedynczej opinii jej polaryzacji (przydzielenie jej do jednej z uprzednio wymienionych grup). Brana jest tu pod uwagę opinia jako całość.
- **Analiza ukierunkowana na cechy produktu** – wyszukanie w opinii poszczególnych aspektów (cech) przedmiotu opinii, a następnie zbadanie stosunku autora wypowiedzi do tego właśnie aspektu. Badana jest nie cała opinia, ale poszczególne jej części odnoszące się do kolejnych cech opisywanego produktu czy usługi.
- **Analiza porównawcza produktów** – badanie opinii na temat jednego produktu, określonej poprzez analizę zdania porównującego go do innego produktu. Konieczne jest zidentyfikowanie w opinii zdań porównujących, a następnie ich analiza ukierunkowana na przedmiot porównania.

Rysunek 1 przedstawia różne text miningowe podejścia do automatycznej analizy opinii konsumenckich. Są one przyporządkowane do poszczególnych działań wyróżnionych w ramach automatycznej analizy opinii konsumenckich.

Koncentrując się na klasyfikacji opinii, można zauważyć, że wszystkie cztery text miningowe podejścia do automatycznej analizy opinii konsumentów znajdują w niej zastosowanie [Lula, Wójcik 2011]:

- **Podejście oparte na słowach (*word-based approach*)** – podstawą tego podejścia jest przekonanie, że znaczenie wypowiedzi (również jej nacechowanie) jest zakodowane w pojedynczych słowach stanowiących dany tekst.

- **Podejście bazujące na wzorcach (*pattern-based approach*)** – w tym podejściu istotne jest przekonanie, że nacechowanie opinii wyznaczają nie pojedyncze słowa, ale zbudowane z nich frazy/związki frazeologiczne. Tak więc konieczne jest wyszukanie wśród słów związków wyrazowych.
- **Podejście bazujące na ontologiach (*ontology-based approach*)** – pojedyncza opinia dotycząca produktu lub usługi może zostać przedstawiona jako instancja ontologii. Następnie instancje te mogą zostać porównane, a na tej podstawie reprezentowane przez nie opinie mogą zostać zaklasyfikowane do jednej z utworzonych grup.
- **Podejście oparte na uczeniu maszynowym (*machine learning approach*)** – dzięki zastosowaniu uczenia maszynowego można zbudować system, który nie tylko na podstawie odpowiednio dobranego uczącego zbioru opinii będzie je klasyfikował do odpowiednich grup, ale również będzie się rozwijał wraz z pojawieniem się nowych, specyficznych opinii.



Rys. 1. Wykorzystanie różnych podejść text miningowych do poszczególnych działań w ramach automatycznej analizy opinii konsumenckich

Źródło: opracowanie własne.

W modelu opartym na uczeniu maszynowym stosowane klasyfikatory wykorzystują odległość pomiędzy opiniami. Jest ona podstawą przypisania opinii pozytywnego lub negatywnego nacechowania.

Jednym ze sposobów klasyfikacji metod pomiaru podobieństwa dokumentów tekstowych jest klasyfikacja ze względu na fakt wykorzystania wiedzy dziedzinowej. Wyróżnić tu można metody niekorzystające z wiedzy dziedzinowej (głównie oparte na macierzy częstości) i te wykorzystujące wiedzę dziedzinową (głównie w postaci ontologii). W niniejszej pracy wykorzystana zostanie miara oparta na macierzy częstości.

3. Materiały i metody badań

Opinie to specyficzny rodzaj danych tekstowych, które mają subiektywny charakter – wyrażają stosunek autora wypowiedzi do przedmiotu opinii. W niektórych serwisach opinie słowne są wspierane oceną punktową lub gwiazdkami. Opinie można podzielić na grupy według ich formatu [Liu 2007]:

- format 1: zalety i wady oraz podsumowanie,
- format 2: zalety i wady,
- format 3: dowolny.

Analiza doniesień literaturowych [Grabowski 2012; Abramowicz 2008] pozwala wyodrębnić dwa główne podejścia stosowane przy ocenie miar podobieństwa dokumentów tekstowych (w tym opinii):

a) podejście polegające na określeniu zbieżności pomiędzy podobieństwem (lub odległością) wyznaczonym za pomocą ocenianej miary a podobieństwem rzeczywistym między dokumentami (określonym na przykład przez człowieka lub wynikającym ze sposobu generowania dokumentów na potrzeby badań symulacyjnych),

b) podejście polegające na analizie własności stosowanych miar poprzez analizę teoretyczną lub symulacyjną.

Przy ocenie zbieżności pomiędzy podobieństwem (lub odległością) wyznaczonym za pomocą ocenianej miary a podobieństwem rzeczywistym pomiędzy dokumentami zauważyć można jedną bardzo istotną trudność. Jest nią dostępność oceny rzeczywistego podobieństwa pomiędzy badanymi tekstami.

Przy dużych zbiorach opinii nieefektywne jest wyznaczanie rzeczywistego podobieństwa badanych tekstów przez badacza lub eksperta w danej dziedzinie. W badaniu przyjęto zatem założenie, że dwie opinie są tym bardziej do siebie podobne, im bardziej zbieżna jest ich ocena punktowa. Stanowi to punkt odniesienia do oceny przydatności miar podobieństwa dokumentów tekstowych, wyznaczając rzeczywiste podobieństwo pomiędzy badanymi tekstami.

Badania empiryczne podzielone zostały na sześć etapów:

1. ekstrakcja opinii,
2. utworzenie macierzy częstości,
3. wygenerowanie macierzy odległości opartej na ocenach użytkowników,
4. wygenerowanie macierzy odległości opartej na macierzy częstości,
5. porównanie dwóch macierzy odległości,
6. użycie algorytmu genetycznego.

W badaniach wykorzystano pakiety tm, RODBC, proxy i genalg języka R oraz aplikację RapidMiner.

3.1. Ekstrakcja opinii i macierz częstości

Opinie wykorzystane w badaniu pochodziły z serwisu Booking.com, pozwalającego na rezerwację pokoi hotelowych i udostępniającego opinie gości o hotelach. Wybrane zostały jedynie opinie w języku polskim. Podzielono je według hoteli, których dotyczyły. Wszystkie opinie w serwisie Booking.com dostępne są w formacie 2 (wady i zalety hotelu). Ponadto do każdej opinii dołączona jest punktowa ocena hotelu.

Pierwszym krokiem w badaniu była ekstrakcja danych z serwisu do bazy danych. W tym celu wykorzystano własny skrypt napisany w języku PHP oraz baza MySQL. Następnie dla każdej opinii polskojęzycznej w bazie danych wygenerowano plik tekstowy zawierający jedynie wady i zalety ocenianego hotelu. W nazwie pliku pojawiała się punktowa ocena. Pliki zostały zgrupowane w folderach zawierających nazwę ocenianego hotelu.

W drugim kroku na podstawie wygenerowanych plików tekstowych utworzono macierz częstości (macierz, której kolumny reprezentują dokumenty, a wiersze − wyrazy). Dla opinii o każdym hotelu tworzone były osobne macierze częstości.

3.2. Macierze podobieństwa i ich porównanie

Kolejnym krokiem w badaniu empirycznym było utworzenie dwóch macierzy odległości. Jedna z nich opierała się na ocenach punktowych. Dla każdej pary opinii policzono różnicę wartości ocen punktowych. Pracę tę wykonano w arkuszu kalkulacyjnym, a następnie w postaci gotowej macierzy odległości wczytano do programu R przy wykorzystaniu pakietu RODBC.

W dalszej części badań na podstawie macierzy częstości utworzono macierz odległości z wykorzystaniem odległości kątowej [Deza i Deza 2009]. W tym celu wykorzystany został pakiet tm języka R. Miarę tę wybrano ze względu na dobre wyniki osiągane przy jej wykorzystaniu we wcześniejszych analizach [Wójcik 2011].

roz

37 observations of 37 variables

	row.names	opinion1(8.8).txt	opinion10(4.6).txt	opinion11(7.9).txt	opinion12(8.8).txt	opinion13(7.1).txt
1	opinion1(8.8).txt	3.986534	3.511385	4.280387	3.986534	4.291590
2	opinion10(4.6).txt	4.569674	3.917934	4.841894	4.569674	4.829189
3	opinion11(7.9).txt	3.459496	2.992120	3.725890	3.459496	3.735925
4	opinion12(8.8).txt	4.188922	3.702500	4.491796	4.188922	4.503185
5	opinion13(7.1).txt	4.094811	3.560351	4.360887	4.094811	4.341064
6	opinion14(6.7).txt	3.379627	2.880257	3.639425	3.379627	3.610230
7	opinion15(6.7).txt	4.135343	3.590580	4.403882	4.135343	4.391938
8	opinion16(5.4).txt	4.186713	3.571949	4.447880	4.186713	4.427411
9	opinion17(7.5).txt	4.388495	3.855891	4.656962	4.388495	4.650412
10	opinion18(7.9).txt	4.126182	3.607475	4.391742	4.126182	4.402065

Rys. 2. Fragment macierzy odległości pomiędzy macierzami odległości opartymi kolejno na ocenach użytkowników i na macierzy częstości dla pojedynczego hotelu

Źródło: opracowanie własne w RStudio.

Po utworzeniu obydwu macierzy odległości (opartej na ocenach użytkowników i na macierzy częstości) zostały one porównane (policzona została odległość pomiędzy nimi). Wykorzystano do tego gotową funkcję *dist* w języku R z pakietu proxy. Rysunek 2 przedstawia fragment wyników tego porównania.

3.3. Algorytm genetyczny

W kolejnym etapie badań wykorzystano algorytm genetyczny [Goldberg 1989], którego zadaniem jest dobór zestawu wag używanych do ważenia macierzy częstości (wagi mogą przyjmować wartości 0 lub 1) tak, aby macierz odległości oparta na niej była jak najbardziej podobna do macierzy odległości opartej na ocenach użytkowników. Minimalizowana jest macierz będąca wynikiem działania funkcji *dist* dla obu macierzy odległości. Funkcja *dist* służy do wyznaczania podobieństwa pomiędzy macierzami.

roz1 ×

37 observations of 37 variables

	row.names	opinion1(8.8).txt	opinion10(4.6).txt	opinion11(7.9).txt	opinion12(8.8).txt	opinion13(7.1).txt
1	opinion1(8.8).txt	0.3056005	1.823175	0.6224823	0.3056005	0.987895
2	opinion10(4.6).txt	2.1615204	1.864381	2.4436036	2.1615204	2.484935
3	opinion11(7.9).txt	0.3056005	1.823175	0.6224823	0.3056005	0.987895
4	opinion12(8.8).txt	0.3056005	1.823175	0.6224823	0.3056005	0.987895
5	opinion13(7.1).txt	0.9275324	1.546567	1.2447393	0.9275324	1.390693
6	opinion14(6.7).txt	0.9327528	1.609092	1.2432068	0.9327528	1.404110
7	opinion15(6.7).txt	0.9457655	1.583337	1.2767175	0.9457655	1.433865
8	opinion16(5.4).txt	1.8764781	1.788525	2.1706261	1.8764781	2.232809
9	opinion17(7.5).txt	0.9327528	1.609092	1.2432068	0.9327528	1.404110
10	opinion18(7.9).txt	0.9121415	1.493627	1.2159915	0.9121415	1.390050

Rys. 3. Fragment macierzy odległości pomiędzy macierzami odległości opartymi kolejno na ocenach użytkowników i na macierzy częstości dla pojedynczego hotelu, powstałej w wyniku działania algorytmu genetycznego

Źródło: opracowanie własne w RStudio.

Wykorzystany w tym celu został wektor wag o liczbie elementów równej liczbie słów w macierzy częstości. Algorytm w 100 powtórzeniach dokonywał zmian w wektorze wag, następnie ważono macierz częstości i porównywano macierze odległości. Rysunek 3 przedstawia fragment macierzy z ostatecznym wynikiem dla jednej z podjętych prób (w każdej z prób uzyskiwano nieznacznie różniące się wyniki).

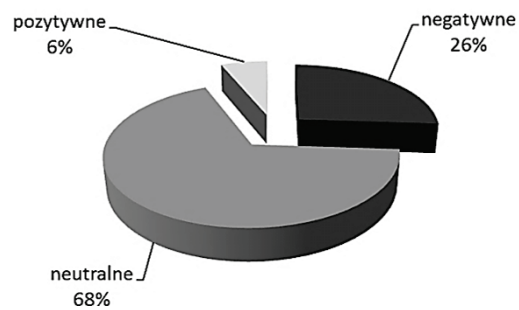
4. Wyniki badań empirycznych

Eksperyment przeprowadzony został na opiniach dotyczących pięciu hoteli. Zbiory opinii liczyły odpowiednio 18, 32, 37, 47 oraz 53 opinie w formie dokumentów tekstowych. Były to wszystkie opinie w języku polskim na temat danych hoteli

dostępne w serwisie Booking.com. Kryterium wyboru hoteli do badań była różnorodność opinii internautów na ich temat oraz ogólna ocena punktowa hoteli. Na jednym ze zbiorów danych badanie zostało wykonane pięciokrotnie. Był to zbiór liczący 37 opinii. Został on wybrany jako najbardziej reprezentatywny.

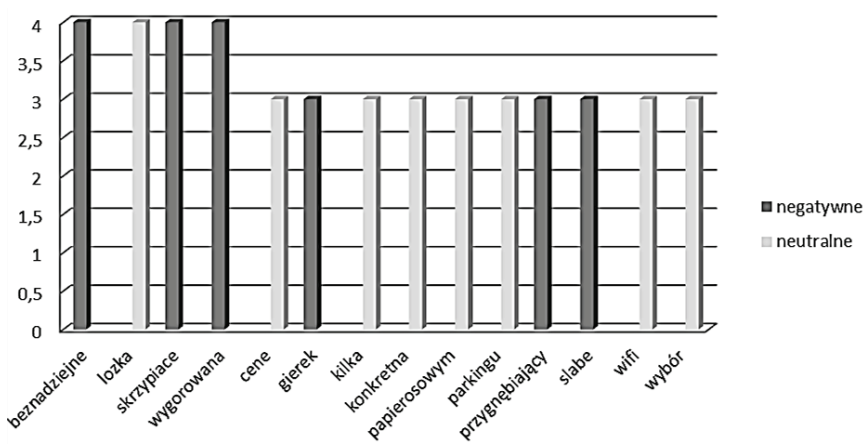
Celem analizy było przetestowanie narzędzia, jakim jest algorytm genetyczny, i ocena jego przydatności w automatycznej analizie opinii konsumenckich. Z punktu widzenia automatycznej analizy opinii konsumenckich ważną kwestią jest nacechowanie słów istotnych.

Zaczynając od zagregowanych wyników z pięciu prób przeprowadzonych dla pojedynczego hotelu na rysunku 4 można zauważyć, że 68% słów istotnych miało nacechowanie neutralne. Pomiędzy pozostałymi słowami widoczna jest dysproporcja na korzyść słów negatywnych (26% w porównaniu z 6% słów pozytywnych). Odzwierciedla to ogólną ocenę danego hotelu przez jego gości.



Rys. 4. Nacechowanie słów istotnych, wybranych w pięciu powtórzeniach badania na zbiorze opinii o pojedynczym hotelu

Źródło: opracowanie własne.



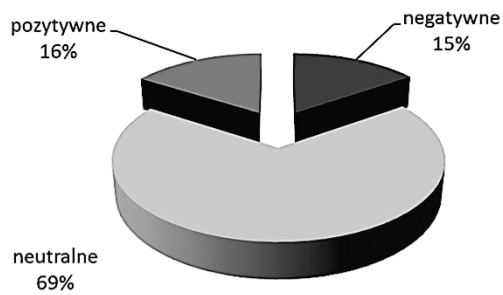
Rys. 5. Słowa powtarzające się w pięciu powtórzeniach badania na zbiorze opinii o pojedynczym hotelu

Źródło: opracowanie własne.

Z kolei rysunek 5 przedstawia słowa istotne, które powtarzały się w kolejnych eksperymentach. Żadne ze słów nie wystąpiło we wszystkich pięciu próbach. Pojawiły się jednak takie, które wystąpiły cztero- lub trzykrotnie. Warto zauważyć, że są to słowa neutralne i negatywne. Zaledwie dwa słowa pozytywne powtórzyły się dwukrotnie (nie zostało to ujęte na wykresie).

Na uwagę zasługuje fakt, iż słowa istotne z punktu widzenia celu analizy nie są słowami występującymi najczęściej w opiniach na temat danego hotelu.

Wykres na rys. 6 przedstawia wyniki pięciu eksperymentów przeprowadzonych na różnych zbiorach danych. Ponownie słowa neutralne stanowią większość wśród słów istotnych (69%). Wartość ta jest podobna do uzyskanej w przypadku badania na jednym zbiorze danych. W tym przypadku można jednak zaobserwować równowagę pomiędzy słowami pozytywnymi i negatywnymi (odpowiednio 16% i 15%) z nieznaczną przewagą tych pierwszych.



Rys. 6. Nacechowanie słów istotnych wybranych w pięciu powtórzeniach badania na zbiorach opinii o różnych hotelach.

Źródło: opracowanie własne.

Dominacja słów o nacechowaniu neutralnym, głównie rzeczowników, wśród słów istotnych sugeruje konieczność powiązania ich z wyrazami pozytywnymi lub negatywnymi, obrazującymi nastawienie autora wypowiedzi do poszczególnych aspektów przedmiotu opinii. Możliwość taką daje wykorzystanie podejścia bazującego na wzorcach, które pozwala na identyfikację w tekście związków frazeologicznych.

5. Podsumowanie

W artykule przedstawione zostały wyniki badań symulacyjnych, dotyczących doboru optymalnego zestawu słów istotnych podczas obliczania podobieństwa (lub odległości) dokumentów tekstowych, stosowanego na potrzeby automatycznej analizy opinii konsumenckich. Zastosowanie algorytmu genetycznego pozwoliło na efektywną selekcję zestawów słów istotnych.

Wśród słów istotnych dominują te o nacechowaniu neutralnym. O przewadze słów pozytywnych nad negatywnymi (lub na odwrót) decyduje sumaryczne nacechowanie opinii na temat danego dobra lub usługi. Ograniczenie liczby słów do tych istotnych nie wpływa na ocenę nacechowania opinii, lecz przyczynia się do zwiększenia efektywności badań poprzez ograniczenie liczby analizowanych danych.

Przeprowadzone analizy pozwalają stwierdzić, iż prowadzenie dalszych badań jest uzasadnione. W przyszłości są więc planowane:

- zwiększenie zbioru badanych opinii,
- zastosowanie podejścia opartego na wzorcach do identyfikacji związków wyrazowych występujących w opiniach,
- próba identyfikacji zestawów słów istotnych dla ogółu opinii na dany temat (np. dla wszystkich opinii o hotelach),
- próby wykorzystania wyników badań w dalszych analizach.

Literatura

Abramowicz W. (2008), *Filtrowanie informacji*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

Deza M.M., Deza E. (2009), *Encyclopedia of distances*, Springer-Verlang, Berlin – Heidelberg.

Goldberg D.E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning. Reading*, Addison-Wesley, Massachusetts.

Grabowski M. (2012), *Naukowa legitymizacja obszaru Systemów Informacyjnych Zarządzania.* Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Liu B. (2007), *Web DataMining. Exploring Hyperlinks, Contents, and Usage Data*, Springer-Verlag, Berlin – Heidelberg.

Lula P., *Automatyczna analiza opinii konsumenckich*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Taksonomia 18, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 53-62.

Lula P., Wójcik K. (2011), *Sentiment analysis of consumer opinions writen in Polish*, „Economics and Management" no 16, s. 1286-1291.

Ohana B., Tierney B. (2009), *Sentiment Classification of Reviews Using SentiWordNet*, IT&T Conference, Dublin Institute of Technology, Dublin.

Pang B., Lee L. (2008), *Opinion Mining and Sentiment Analysis*, „Foundations and Trends in Information Retrieval" 2(1-2), s. 1-135.

Wójcik K. (2011), *Analiza porównawcza miar podobieństwa tekstów*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Taksonomia 18, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 340-348.

Wikipedia. http://pl.wikipedia.org/wiki (12 października 2012).

SELECTION OF THE OPTIMAL SET OF RELEVANT WORDS IN CONSUMERS OPINIONS IN THE CONTEXT OF THE OPINION MINING

Summary: Sentiment analysis is the research area that can have a significant impact on today's business. A lot of sentiment analysis studies base on similarity (or distance) between opinions as special kinds of texts. The problem that must be solved is to determine the best similarity measure to deal with opinions. The main objective of this paper is to conduct the analysis of the selection of optimal set of words relevant when calculating the similarity (distance) of text documents in the context of the opinion mining. Computations are conducted in R language, RapidMiner application and spreadsheet.

Keywords: text-mining, Web-mining, taxonomy, text document classification, evaluation of measures.