METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XV/4, 2014, str. 101 – 110

ZASTOSOWANIE TECHNIK EKSPLORACJI TEKSTU DO ANALIZY OPINII KONSUMENCKICH

Marcin Olszewski, Tomasz Ząbkowski
Katedra Informatyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: marcin_olszewski@sggw.pl, tomasz_zabkowski@sggw.pl

Streszczenie: W niniejszej publikacji zaproponowano jedną z metod eksploracji danych – reguły asocjacyjne do wykrycia zależności w opiniach konsumenckich, na przykładzie opinii jednego z hoteli amerykańskich. Wykorzystanie tej techniki wynikało m.in. z dużej ilości dostępnych danych oraz faktu, że otrzymane reguły w sposób niezwykle czytelny prezentują zależności znalezione w danych. W badaniu odkryto szereg reguł, które mogą stanowić cenne źródło informacji o jakości usług oraz postrzeganiu obiektu przez klientów korzystających z usług hotelowych.

Słowa kluczowe: eksploracja tekstu, reguły asocjacyjne, opinie konsumenckie

WPROWADZENIE I OKREŚLENIE CELU BADANIA

Celem prezentowanych badań było zastosowanie reguł asocjacyjnych do analizy opinii konsumenckich. Tego typu analiza jest narzędziem, które może dostarczyć istotnych informacji mających wpływ na wizerunek firmy, co jest nie bez znaczenia dla firm działających na bardzo konkurencyjnym rynku. Jednakże, pozyskiwanie takich informacji nie jest procesem łatwym i konieczne w tym celu jest przetworzenie ogromnej liczby dokumentów takich jak ankiety, kwestionariusze, opinie. Problem ten wymaga nieco innego spojrzenia na dane i doboru określonego sposobu ich analizowania za pomocą technik eksploracji danych [Larose 2006, Han i Kamber 2001]. Metody te mają kluczowe znaczenie w przypadku badania opinii konsumenckich, umożliwiają one bowiem tworzenie reprezentacji tekstu, która będzie posiadała cechy tekstu wejściowego, a przy tym formę adekwatną do wykorzystania przez algorytmy. Dokumenty tekstowe można traktować jako zbiory z atrybutami o charakterze nominalnym, dlatego też szczególnie użytecznym sposobem ich analizowania może być analiza koszykowa

[Pasztyła 2005], która za pomocą reguł wskazuje na współwystępujące słowa [Amir i in. 2005]. Znalezienie takich zestawów słów wewnątrz opinii konsumenckich, przy uwzględnieniu ich nacechowania semantycznego pozwala na stworzenie bazy wiedzy, w oparciu o którą możliwe jest także dokonywanie klasyfikacji nowych opinii. W wymienionej literaturze przedmiotu opisuje się szeroko zastosowanie analizy koszykowej do wyszukiwania reguł w różnych bazach, jednak w przypadku baz nieustrukturyzowanych, takich jak bazy tekstowe, nie przeprowadzono jak dotąd eksperymentów, które dodatkowo uwzględniałyby rozpatrywane w pracy aspekty dotyczące przetwarzania tekstów pod kątem usprawnienia działania algorytmów asocjacyjnych. To właśnie jest głównym zamierzeniem autorów niniejszego artykułu i stanowi o jego wartości.

METODY ZMNIEJSZANIA REPREZENTACJI TEKSTU OPINII KLIENCKICH

Każda opinia konsumencka jest dokumentem tekstowym, który nie posiada konkretnej struktury. Brak sformalizowanej postaci takiego dokumentu uniemożliwia jego natychmiastową klasyfikację i tym bardziej pozyskanie z niego określonych informacji. Na tym etapie niezbędne jest wstępne przetworzenie opinii, czego efektem będzie odpowiednia dla danego algorytmu postać tekstu, zawierająca ponadto zmniejszoną reprezentację. Możliwości algorytmów eksploracji tekstu są mocno ograniczone jeżeli chodzi o pracę na dużej ilości danych (duża złożoność obliczeniowa i długi czas pracy), dlatego etap ten obejmuje przekształcenie tekstu do zmniejszonej i uproszczonej postaci. Postać taka umożliwia o wiele szybszą i bardziej efektywną analizę danych. W celu zmniejszenia wielkości wejściowych danych tekstowych najczęściej stosuje się podejście tzw. „worka słów" (ang. bag of words), w którym ważniejsza jest częstość występowania słów w poszczególnych dokumentach niż miejsce, czy kolejność ich wystąpienia. Zasadniczo, większość z metod pozyskiwania zmniejszonych reprezentacji tekstu opiera się na prostych statystykach lub zależnościach. Niżej wymieniono metody zmniejszania reprezentacji dokumentów tekstowych, które są istotne dla efektywnego przetwarzania tekstu.

Metoda oparta na prawie Zipf'a

Zgodnie z teorią informacji w każdym języku naturalnym istnieje zależność, która mówi, że rozkład częstości słów występujących w danym języku nie jest losowy, czyli taki, w którym wybrane słowo X występuje z tą samą częstością co słowo Y. Słów najczęściej używanych jest proporcjonalnie wiele razy więcej niż słów najrzadziej używanych w dowolnych rzeczywistym tekście. Ten nierównomierny rozkład słów w językach naturalnych został potwierdzony przez badanie amerykańskiego lingwisty George'a Zipf'a. Prawo to umożliwia odnalezienie zależności w ogromnych ilościach danych tekstowych, które na pierwszy rzut oka mogą wydawać się jednolite. Ponadto prawo to można

wykorzystać do określenia rangi słów. Gdy każdemu słowu z rozkładu Zipf'a przypisze się wartość oznaczającą pozycję w rankingu ważności takiego słowa na podstawie częstości jego wystąpienia, to częstość występowania słów będzie odwrotnie proporcjonalna do pozycji tego słowa w rankingu ważności słów [Ward 1997].

Metoda oparta na stop liście słów

Oprócz własności, którą zauważył Zipf, każdy język charakteryzuje specyficzna konstrukcja posiadająca odpowiednie kryteria składniowe i fleksyjne. Do budowy zdań używa się różnych części mowy i są to (w zależności od języka): zaimki, przyimki, rodzajniki, spójniki, wykrzykniki. Słowa należące do wymienionych kategorii mają bardzo wysoką częstość wystąpień, ale nie niosą żadnej użytecznej wiedzy. Metoda stop listy [Rajaraman i Ullman 2012] (ang. stop list, stop word) polega na pominięciu tych słów na początkowym etapie przygotowania danych w celu usprawnienia pracy algorytmu.

Metoda przycinania

Poza ograniczaniem liczności zbioru słów poprzez tworzenie stop listy, można także zmniejszać reprezentację tekstu za pomocą miar statystycznych – jest to tzw. przycinanie (ang. pruning). Rozwiązanie takie polega na usuwaniu słów najczęściej lub zbyt często występujących w danym dokumencie tekstowym oraz słów, których częstość występowania jest bardzo mała. Określenie progów oddzielających słowa nieistotne z powodu zbyt dużej lub zbyt małej częstości użycia, znacznie zmniejsza rozmiar reprezentacji, dzięki czemu poprawia się efektywność przetwarzania danych oraz redukuje szum informacyjny, nie zmieniając przy tym znacząco wyników działania algorytmu eksploracji tekstu.

ZASTOSOWANE TECHNIKI – REGUŁY ASOCJACYJNE

Wyodrębnienie występujących zestawów słów w treściach opinii może zostać przeprowadzone w oparciu o analizę koszykową. Jej celem jest wyszukiwanie zależności między badanymi obiektami – w tym przypadku między poszczególnymi słowami – i przedstawienia ich w postaci reguł. Reguły te nazywa się regułami asocjacyjnymi i mają one postać:

$$[poprzednik: warunek] \Longrightarrow [następnik: warunek].$$

Wyszukiwanie reguł asocjacyjnych w opiniach sprowadza się do ilościowego określenia relacji pomiędzy poszczególnymi słowami. Ocenę przydatności reguł asocjacyjnych wyznacza się miarami jakości [Pasztyła 2005], [Tan i in. 2005]: wsparciem (ang. support), które mówi jak często otrzymana reguła występuje całej bazie opinii, pewnością (ang. confidence), która mówi, jak silna jest reguła, co odpowiada prawdopodobieństwu warunkowemu, że jeżeli w losowej opinii, pojawiło się słowo X to pojawi się również słowo Y, oraz przyrostem (ang. lift), który mówi jaki wpływ ma pojawienie się w opinii słowa X na wystąpienie w opinii

również słowa Y. Wymienione miary są obiektywnym sposobem na ocenę istotności i siły niesionych przez reguły informacji, a ich formułowanie jest istotne, gdyż często dla określonych danych wejściowych z wieloma atrybutami powstaje bardzo dużo potencjalnych reguł, z których nie wszystkie są użyteczne.

ANALIZA OPINII KONSUMENCKICH

Przygotowanie danych

Przeprowadzony eksperyment polegał na stworzeniu zbioru reguł asocjacyjnych dla przykładowego dokumentu z opiniami na temat jednego z amerykańskich hoteli. Analizę koszykową przeprowadzono algorytmem Apriori [Agrawal i Srikant 1994]. Algorytm ten w każdym kroku iteracji generuje zbiory kandydujące o liczności większej niż w poprzednim kroku, które po spełnieniu określonego progu minimalnego wsparcia stają się zbiorami częstymi. Ze zbiorów częstych wybiera się te, które osiągają wymagany poziom ufności i z nich tworzy się reguły asocjacyjne. Dane ujęte w badaniu pochodzą z repozytorium internetowego, które opracowali Ganesan i Zhai [Ganesan i Zhai 2011]. Cały zbiór danych zawiera listę rzeczywistych opinii hoteli z 10 różnych miast świata. Analiza została ograniczona do przebadania opinii jednego z hoteli, mianowicie hotelu Riviera z Las Vegas. Hotel ten posiada 576 opinii, w których łączna liczba użytych słów wynosi 123727. Treść przykładowych opinii prezentuje Tabela 1. Do analizy wykorzystano treść opinii, gdzie każda opinia to jeden unikalny koszyk, czyli zbiór słów składający się na tę opinię (transakcję). Zanim możliwe było wyszukanie powiązań, dane należało sprowadzić do formatu transakcyjnego.

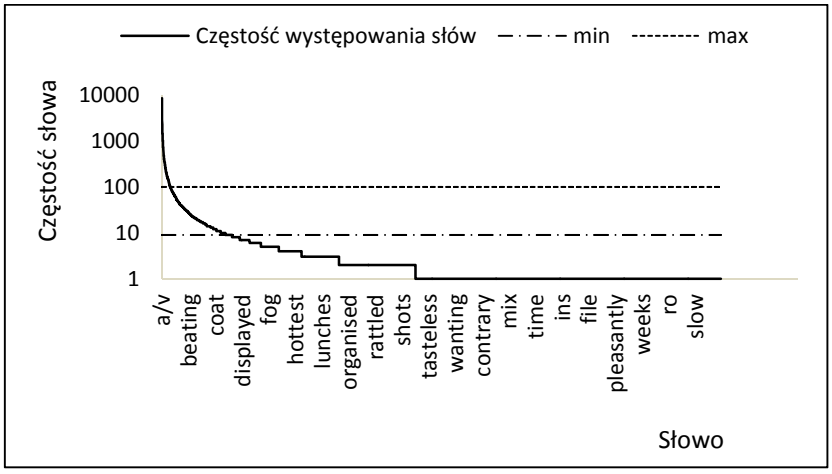
Tabela 1. Przykładowe opinie na temat hotelu

Nr opinii	Treść
1	Hotel Room: Monaco Tower, 4927, king bed. Bed was comfortable and pillows had some support. It won't remind you of home but a solid nights sleep is easily achievable. Room I'd guess is 300 sq ft. Dark brown decor and white sheets. Flat screen tv, 26-32 inches, somewhere around that size. Three dresser drawer, two night stands, desk and an arm chair with foot rest. Safe (I think you have to pay for it) in night stand. View of the Hilton and parking deck behind the hotel. Con: walls seemed to be thin or one of our neighbors was deaf. Wed night we could hear one of our neighbors Tv's like it was in our room.
2	The location is a bit far north of the main action on the strip, and still a long bus ride to Fremont Street. The room was small and below average. The staff are rude, and the casino staff even worse. Find a better place to stay.

Źródło: opracowanie własne

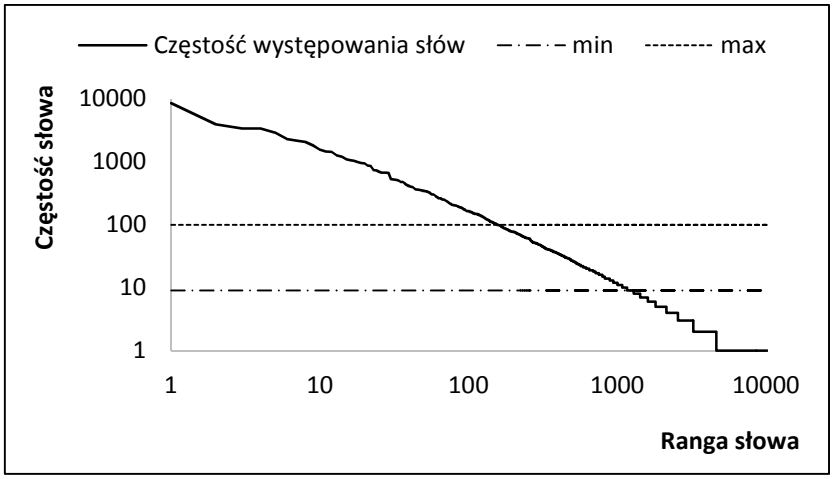
Z każdej opinii usunięto wszelkie znaki interpunkcyjne (kropki, przecinki, średniki, myślniki, znaki zapytania, wykrzykniki, dwukropki, nawiasy). Powstała tabela zawiera tylko słowa języka angielskiego (oraz ewentualne nieznane wyrażenia wynikające z błędów osób piszących opinię). Szereg słów nie niesie żadnej użytecznej informacji, dlatego zawartość tabeli została poddana zabiegom zmniejszającym liczbę słów we wszystkich opiniach. W tym celu została stworzona stop lista, która zawiera słowa będące zaimkami (31 sztuk), przyimkami (72 sztuk), rodzajnikami (5 sztuk), spójnikami (34 sztuk) i wykrzyknikami (114 sztuk) [Słownik spójników 2013], [Słownik przyimków 2013], [Słownik wykrzykników 2013]. W badaniu uwzględniono także wielkość liter dla słów rozpoczynających zdania. Ten zabieg pozwolił zmniejszyć reprezentację tekstu usuwając łącznie 48002 słów (powtarzających się wystąpień słów ze stop listy), co spowodowało, że w bazie pozostało 75725 słów. W następnym kroku usunięte zostały słowa zbędne w oparciu o metodę przycinania, polegającą na usuwaniu słów o zbyt dużej i zbyt małej częstości wystąpień [Borycki i Sołdacki 2002]. Progi te zostały dobrane eksperymentalnie, w taki sposób, aby w końcowej ocenie miary statystyczne dla reguł asocjacyjnych pozostawały możliwe najwyższe tj. najlepsze w rozpatrywanym przypadku. Zostały one pokazane na wykresach nr 1 i nr 2. Słowa o częstości większej niż 100 to głównie słowa ze stop listy, ale także słowa, które nie niosą żadnej użytecznej informacji. Słowa o częstości niższej niż 9, to najczęściej błędy użytkowników lub słowa tak rzadkie, że usunięcie ich nie wpłynie na wyniki, ale poprawi efektywność przetwarzania, gdyż tych słów jest aż 8918 (17324 słów uwzględniając ich powtórzenia). Schodkowość wykresów wynika z powtarzających się liczności poszczególnych słów.

Rysunek 1. Wykres zależności poszczególnych słów z opinii o hotelu do częstości ich wystąpień (z uwagi na ograniczenia prezentowane są tylko wybrane słowa)



Źródło: opracowanie własne

Rysunek 2. Wykres zależności rang poszczególnych słów z opinii o hotelu do częstości ich wystąpień



Źródło: opracowanie własne

Dla tak dobranych progów zostało pominiętych 9074 unikatowych słów (uwzględniając stop listę liczba ta zmniejszy się z uwagi na powtarzające się słowa). Metoda przycinania pozwoliła usunąć łącznie 48940 słów, co spowoduje, że w bazie pozostanie 26785 słów. Wewnątrz wszystkich opinii są jednak słowa powtarzające się, gdyż w jednej opinii możliwe było użycie kilku takich samych słów. Do analizy wystarczyła reprezentacja składająca się z unikatowych słów wewnątrz opinii [Morzy 2013]. Po wykluczeniu powtarzających się słów wewnątrz opinii pozostało 23609, co stanowiło zbiór wejściowy do analizy. Tabela 1 prezentuje transakcyjną postać opinii klientów hotelu po przygotowaniu.

Tabela 2. Transakcyjna postać opinii wg przygotowania

Nr opinii	**Słowo**
1	always
1	anyway
1	ask
⋮	⋮
2	action
2	average
2	far
⋮	⋮
3	bathroom
3	best
⋮	⋮

Źródło: opracowanie własne

Wykrywanie reguł asocjacyjnych w tekstach opinii

W badaniu rozważano reguły dwu i trzyelementowe. Ustalony został także minimalny poziom wsparcia na poziomie 3% oraz pewności reguły na poziomie 30%. Dla tak skonfigurowanego procesu znaleziono 68 reguł. Wyniki dla dziesięciu pierwszych reguł przedstawiono w tabelach 3, 4, 5 uwzględniając szeregowanie według miar wsparcia, pewności i przyrostu.

Na podstawie przeprowadzonej analizy najczęstszymi zwrotami (popartymi pewnością) użytymi w opiniach użytkowników są m.in. pary: reading–reviews, slots–fun, shop–coffee, screen–flat, flat–tv, monte–carlo, minutes–took, line–long, checkin–took. U wielu oceniających padła formuła zawierająca słowa reviews i reading. Może to oznaczać, że użytkownicy chętnie czytają i dzielą się opiniami na temat hotelu, co wskazuje, że ten kanał komunikacji jest niezwykle opiniotwórczy dla wielu użytkowników, w tym także dla potencjalnych klientów.

Tabela 3. Wybrane reguły asocjacyjne dla tabeli słów (posortowane wg wsparcia)

Przyrost	Wsparcie (%)	Pewność(%)	Reguła
5.67	5.73	84.62	reading => reviews
16.94	5.73	100.00	monte => carlo
3.92	5.38	47.69	slots => fun
5.57	4.86	60.87	shop => coffee
13.37	4.51	81.25	screen => flat
5.02	4.34	71.43	flat => tv
3.36	4.34	45.45	minutes => took
2.49	4.34	33.78	checkin => took
3.00	4.17	37.50	line => long
2.24	3.99	30.67	use => bit

Źródło: opracowanie własne

Tabela 4. Wybrane reguły asocjacyjne dla tabeli słów (posortowane wg pewności)

Przyrost	Wsparcie (%)	Pewność(%)	Reguła
16.94	5.73	100.00	monte => carlo
14.31	3.47	86.96	tv & screen => flat
5.67	5.73	84.62	reading => reviews
13.37	4.51	81.25	screen => flat
14.40	3.47	80.00	tv & flat => screen
5.40	3.47	76.92	screen & flat => tv
13.37	4.51	74.29	flat => screen
5.05	3.99	71.88	screen => tv
5.02	4.34	71.43	flat => tv
4.25	4.51	63.41	read => reviews

Źródło: opracowanie własne

Tabela 5. Wybrane reguły dla tabeli słów (posortowane wg przyrostu)

Przyrost	Wsparcie (%)	Pewność(%)	Reguła
16.94	5.73	100.00	monte => carlo
14.40	3.47	80.00	tv & flat => screen
14.31	3.47	86.96	tv & screen => flat
13.37	4.51	81.25	screen => flat
5.67	5.73	84.62	reading => reviews
5.57	4.86	60.87	shop => coffee
5.40	3.47	76.92	screen & flat => tv
5.02	4.34	71.43	flat => tv
3.36	4.34	45.45	minutes => took
3.00	4.17	37.50	line => long

Źródło: opracowanie własne

Współwystępujące pary słów: line–long, minutes–took, checkin–took mogą wskazywać na długi czas oczekiwania w kolejce na zameldowanie lub inne czynności hotelowe, co spowodowało, że wielu użytkowników nie wyraziło się pozytywnie o hotelu. Z kolei zdania zawierające pary słów: shop–coffee, flat–tv, slots–fun mogą świadczyć o zadowoleniu użytkowników z wyposażenia pokoi hotelowych oraz dostępności sklepu, kawiarni czy rozrywki, co zaowocowało pozytywnymi opiniami. Natomiast reguła monte–carlo, nie jest użyteczna, gdyż słowa te stanowią nazwę własną oznaczającą nazwę części budynku hotelowego.

WNIOSKI

Na podstawie przeprowadzonych badań można sformułować następujące wnioski.

Reguły asocjacyjne znajdują zastosowanie do odnajdywania zależności w danych tekstowych. Dane te jednak zazwyczaj mają obszerną reprezentację i nieustrukturyzowaną formę, co powoduje, że wymagają one odpowiedniego przygotowania. Opierając się na prawie Zipf'a, metodzie stop listy oraz przycinaniu, można zmniejszyć reprezentację tekstu, co jest konieczne, aby zastosować wybrany algorytm reguł asocjacyjnych. Analiza taka może być pomocna do odpowiedzi na pytanie jak postrzegany jest przez klientów dany obiekt hotelowy. Otrzymane wyniki analizy tekstu sugerują, że pary słów m.in.: long–line, took–minutes, checkin–took reprezentują negatywną klasę opinii. Pokazuje to, że w tym obszarze działalności hotelu konieczne są usprawnienia, które przyczynią się do podniesienia jakości obsługi. Z kolei teksty zawierające m.in. frazy: shop–coffee, tv–flat, slots–fun mogą posłużyć jako klasyfikator korzystnych dla hotelu opinii.

Przytoczone powyżej reguły stanowią cenne źródło informacji o opiniach i nastrojach klientów korzystających z usług hotelowych. Nie bez znaczenia pozostaje fakt, że otrzymane reguły w sposób niezwykle przystępny ujmują najważniejsze prawidłowości w bardzo dużym zbiorze opinii.

BIBLIOGRAFIA

Agrawal R., Srikant R. (1994) Fast Algorithms for Mining Association Rules, IBM Research Report RJ9839, IBM Almaden Research Center San Jose, California.

Amir A., Auman Y., Feldman R., Fresko M. (2005) Maximal Association Rules: A Tool for Mining Associations in Text, Journal of Int. Information Systems, str. 333–345.

Borycki Ł., Sołdacki P. (2002) Automatyczna klasyfikacja tekstów, in: Mat. III Krajowej Konferencji: Multimedialne i Sieciowe Systemy Informacyjne, str. 473-481.

Ganesan K., Zhai C. (2011) Opinion-Based Entity Ranking, Information Retrieval. [Online http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset]

Gunther R.,Levitin L., Shapiro B., Wagner P. (1996) Zipf's law and the effect of ranking on probability distributions, Int. Journal of Theoretical Physics, 35(2), str. 395-417.

Han J., Kamber M. (2001) Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher.

Larose D. (2006) Odkrywanie wiedzy z danych, Wydawnictwo Naukowe PWN, Warszawa.

Morzy, T. (2013). Eksploracja danych. Wydawnictwo Naukowe PWN, Warszawa.

Pasztyła A. (2005) Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej, Data mining: poznaj siebie i swoich klientów (pub. elektr.). Statsoft, Kraków.

Rajaraman A., Ullman J.D. (2012) Data Mining. Mining of Massive Datasets. Cambridge University Press, New York.

Słownik przyimków w języku angielskim (2013) [Online http://www.englishclub.com/grammar/prepositions-list.htm (2013-12-30)]

Słownik spójników w języku angielskim (2013) [Online http://www.english-grammar-revolution.com/list-of-conjunctions.html (2013-12-30)]

Słownik wykrzykników w języku angielskim (2013) [Online http://www.vidarholen.net/contents/interjections (2013-12-30)]

Tan P., Steinbach M., Kumar V. (2005) Introduction to Data Mining, Addison-Wesley, Boston.

Ward M. (1997) 50 najważniejszych problemów zarządzania, Wydawnictwo Profesjonalnej Szkoły Biznesu, Kraków.

APPLICATION OF TEXT MINING TECHNIQUES FOR THE CUSTOMER REVIEWS ANALYSIS

Abstract: This paper presents application of one of data mining techniques – association rules to analyze customer reviews, based on the data gathered at one of the American hotels. The application of association rules is due to the large volume of available review data and the fact that the rules can be presented in a very clear and meaningful way. The study resulted in a number of interesting rules that can be a valuable source of information about the quality of services and the perception of the hotel by the clients.

Keywords: text mining, association rules, customer reviews