PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 — 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Paweł Lula
Akademia Ekonomiczna w Krakowie

WYKORZYSTANIE INFORMACJI POCHODZĄCYCH Z DOKUMENTÓW TEKSTOWYCH W PROBLEMACH MODELOWANIA I KLASYFIKACJI

1. Wstęp

Wśród zasobów informacyjnych gromadzonych i przetwarzanych przez człowieka znaczny udział mają dokumenty tekstowe. Zawarte w nich informacje mogą w wielu przypadkach wpłynąć pozytywnie na proces modelowania, prognozowania i klasyfikacji rozpatrywanych zjawisk. Uwzględnienie w procesie modelowania informacji pochodzących z dokumentów tekstowych ma charakter procedury dwuetapowej. Pierwszym krokiem jest pozyskanie informacji z dokumentu i nadanie jej ściśle zdefiniowanej postaci. Drugi etap polega na wykorzystaniu ustrukturyzowanych informacji jako danych wejściowych konstruowanego modelu prognostycznego. Realizacja przedstawionego zagadnienia należy do podstawowych zadań text miningu.

Niniejszą pracę rozpoczyna prezentacja podstawowych zagadnień związanych z text miningiem. Przedstawiona zostanie jego definicja, związki z innymi metodami badań oraz krótki przegląd problemów badawczych rozpatrywanych na jego gruncie. Kolejny punkt pracy dotyczy metod pozyskiwania informacji z dokumentów tekstowych i sposobów jej reprezentacji. Kontynuacją tego zagadnienia są rozważania zawarte w następnej części pracy, które dotyczą sposobów wykorzystania pozyskanych informacji w procesie analizy. Poruszone w artykule problemy zilustrowane zostaną przykładem wykorzystującym rzeczywiste dane tekstowe. Końcowa część pracy zawiera wnioski i spis wykorzystanej literatury.

2. Ogólna charakterystyka text miningu

Text mining jest próbą automatyzacji zadań związanych z przetwarzaniem informacji o charakterze tekstowym, realizowaną za pomocą metod analizy danych. W

obecnych czasach, gdy zalew informacyjny jest zjawiskiem powszechnym, potrzeba wsparcia człowieka w procesie analizy tekstowych zasobów informacyjnych jest nie do przecenienia. Jednakże jest to zadanie trudne w realizacji. Za główne przyczyny problemów należy uznać niski stopień ustrukturyzowania tekstów oraz wieloznaczność tej formy przekazu i uzależnienie sposobu interpretacji od kontekstu.

Podejście text miningowe ma stosunkowo krótką historię. Za moment jego powstania uznaje się rok 1999, kiedy to Marti A. Hearst opublikowała pracę, w której zdefiniowała text mining jako proces mający na celu wydobycie z zasobów tekstowych nieznanych wcześniej informacji [Hearst 1999]. Kolejne lata stały się okresem dynamicznego rozwoju nowego sposobu badań. Na potrzeby analizy danych tekstowych zaadaptowane zostały osiągnięcia i metody statystyki, informatyki, data miningu, uczenia maszynowego oraz sztucznej inteligencji (zwłaszcza rozwiązania dotyczące przetwarzania języka naturalnego i wyszukiwania oraz reprezentacji informacji). Ten bogaty zestaw metod może służyć do realizacji różnorodnych zadań, do których zalicza się przede wszystkim:

- pozyskiwanie informacji z tekstów,
- identyfikację wiadomości zawierających określone treści,
- generowanie streszczeń,
- klasyfikację wzorcową dokumentów,
- klasyfikację bezwzorcową dokumentów (analizę skupień),
- identyfikację i wizualizację powiązań pomiędzy dokumentami,
- generowanie odpowiedzi na pytania zadawane w języku naturalnym przez użytkownika systemu.

Sposób realizacji każdego z wymienionych powyżej zadań jest inny, ale ogólny schemat analizy jest w każdym przypadku zbliżony i obejmuje następujące etapy:

- określenie celu, zakresu i kosztów badań,
- wstępne przetworzenie dokumentów,
- określenie sposobu reprezentacji informacji zawartych w dokumentach i jej pozyskanie oraz ustrukturyzowanie,
- konstrukcję modelu,
- ocenę jakości modelu,
- zastosowanie skonstruowanego modelu i interpretację uzyskanych wyników.

Każdy z przedstawionych powyżej etapów należy uznać za bardzo ważny, ale z punktu widzenia problematyki zasygnalizowanej przez tytuł niniejszego opracowania szczególna uwaga poświęcona zostanie zadaniom wymienionym w punkcie trzecim i czwartym. Wymienione tam zadania są mocno ze sobą powiązane, gdyż przyjęcie określonego sposobu reprezentacji informacji pozyskanej z dokumentów tekstowych determinuje sposób jej dalszej analizy, a więc również kwestie dopuszczalnych typów modeli, ich szacowania i oceny.

3. Pozyskiwanie i reprezentacja informacji z dokumentów tekstowych

Zarówno sposób pozyskania informacji, jak i sposób jej dalszego przetworzenia jest zdeterminowany przez przyjęty sposób reprezentacji informacji pozyskanych z dokumentów. Do najczęściej spotykanych metod należy zaliczyć:

- reprezentacje bazujące na macierzy częstości,
- reprezentacje listowe,
- reprezentacje drzewiaste,
- reprezentacje grafowe.

3.1. Reprezentacje bazujące na macierzy częstości

Idea tego sposobu przechowywania informacji jest bardzo prosta. Mianowicie tworzona jest macierz zawierająca dane dotyczące liczby wystąpień każdego słowa w każdym z rozpatrywanych dokumentów. Przy założeniu, że rozpatrywany jest zbiór m dokumentów zawierający n równych słów, to macierz częstości **X** będzie macierzą o m kolumnach i n wierszach, w której element x_{ij} będzie informował o liczbie wystąpień i-tego wyrazu w j-tym dokumencie.

Do podstawowych zalet reprezentacji macierzowej należy zaliczyć bezproblemowy sposób wyznaczania macierzy częstości, prostotę interpretacji jej elementów oraz łatwość jej dalszego przetwarzania. Zasadniczą wadą jest natomiast pominięcie kolejności informacji wynikających z kolejności występowania wyrazów w przetwarzanych dokumentach.

Przedstawione powyżej zalety sprawiają, że macierz częstości (pomimo niewątpliwych jej wad) jest najczęściej wykorzystywaną metodą reprezentacji informacji pozyskanych z dokumentów tekstowych.

3.2. Reprezentacje listowe

Lista jest strukturą danych grupującą elementy uporządkowane liniowo. W kontekście analizy zawartości dokumentów tekstowych listy będą tworzone z wyrazów składających się na przetwarzany dokument, przy czym punktem wyjścia może być pierwotny dokument lub też dokument po wstępnym przetworzeniu. W zależności od celu badań możliwe jest utworzenie jednej listy reprezentującej cały dokument lub wielu list odpowiadających poszczególnym fragmentom dokumentu (na przykład rozdziałom, akapitom, zdaniom, frazom).

Porównując reprezentację wykorzystującą macierz częstości z reprezentacją listową, należy wskazać, że przewagą tej drugiej jest zachowanie informacji o kolejności wyrazów. Jednakże zastosowanie struktur listowych znacznie utrudnia prowadzenie dalszych analiz, gdyż liczba algorytmów obliczeniowych przystosowanych do działań na listach jest znacznie mniejsza od liczby algorytmów przystosowanych do działań na danych w postaci macierzy.

3.3. Reprezentacje drzewiaste

Do najistotniejszych elementów zawartych w dokumentach tekstowych należy zaliczyć opis faktów i istniejących pomiędzy nimi związków. W celu dokonania standaryzacji opisu tego typu elementów tworzone są ontologie będące zbiorem sformalizowanych pojęć służących do reprezentacji wybranego fragmentu rzeczywistości [Gruber 1993]. W wyniku zastosowania ontologii obiekty opisane w dokumencie przedstawiane są w postaci drzew, do których reprezentacji stosowany jest powszechnie język XML.

W porównaniu z ujęciami opisanymi we wcześniejszych podpunktach przekształcenie dokumentu do postaci drzewa (drzew) nie jest zadaniem prostym. Aby je zrealizować, należy zdefiniować zbiór wzorców, np. w postaci wyrażeń regularnych, pozwalających na zidentyfikowanie istotnych treści, lub też należy przeprowadzić manualnie całą procedurę konwersji i wykorzystać dokumenty źródłowe wraz z informacją o sposobie ich zaklasyfikowania w charakterze zbioru uczącego dla samouczącej się procedury konwertującej.

3.4. Reprezentacje grafowe

Najbardziej zaawansowaną strukturą służącą do reprezentacji informacji pochodzących z dokumentu są grafy. Popularność reprezentacji grafowej jest związana z upowszechnianiem się koncepcji sieci semantycznej (Semantic Web), której twórcą jest Tim Berners-Lee. Sieć semantyczna jest uniwersalnym narzędziem reprezentacji wiedzy. Przyjmuje postać grafu przechowującego informacje o obiektach i zachodzących pomiędzy nimi relacjami. Wraz z rozwojem metod reprezentacji wiedzy prowadzone są prace nad narzędziami służącymi do jej wykorzystania ([Berners-Lee 1998]; http://en.wikipedia.org/wiki/Semantic_web).

Przekształcenie dokumentu tekstowego do postaci sieci semantycznej jest zadaniem trudnym do automatyzacji. Z tego powodu nie należy się spodziewać masowej migracji istniejących dokumentów do tej postaci. Panuje jednak powszechne przekonanie, że stosowany obecnie w internecie system WWW w niedługim czasie ewoluować będzie w kierunku sieci semantycznej, gdyż tylko tego typu zmiany mogą w sposób istotny podnieść stopień wykorzystania zasobów internetowych.

4. Konstrukcja modelu

Ten etap prac badawczych należy rozpocząć od dokonania wyboru rodzaju modelu. Podejmując decyzję, należy uwzględnić przede wszystkim przyjęty sposób reprezentacji informacji oraz rodzaj rozwiązywanego problemu badawczego.

Najmniej problemów stwarza reprezentacja macierzowa. Przy jej zastosowaniu informacje zawarte w poszczególnych dokumentach reprezentowane są przez wek-

tory będące kolejnymi kolumnami macierzy **X**, dzięki czemu można w prosty sposób obliczyć miary podobieństwa lub odległości pomiędzy dokumentami. Istotną zaletą reprezentacji macierzowej jest również możliwość przetwarzania tak reprezentowanych informacji za pomocą klasycznych metod analizy danych.

W przypadku zastosowania reprezentacji listowej wykonywanie analiz jest utrudnione, co uwidacznia się nawet przy próbie wyznaczenia odległości pomiędzy fragmentami tekstu. Najpopularniejszą propozycją rozwiązania tego zagadnienia jest zastosowanie odległości edycyjnej (odległości Levenshteina), określającej liczbę operacji elementarnych potrzebnych do przekształcenia pierwszej listy w drugą.

Poważne problemy mogą się również pojawić przy próbie zastosowania reprezentacji drzewiastej lub grafowej. Przeprowadzenie badań wymagających określenia podobieństwa pomiędzy reprezentowanymi w ten sposób dokumentami stwarza konieczność wyznaczenia i właściwego zagregowania podobieństwa (lub odległości) zarówno w zakresie struktury, jak i w zakresie semantyki. Problematyka ta jest często poruszana w literaturze (np. [Montes-y-Gómez i in. 2000; Nierman i in. 2002; Oldakowski, Bizer 2005]). Analizy wykraczające poza wyznaczanie podobieństwa lub odległości pomiędzy tekstami reprezentowanymi w postaci drzew lub grafów przeprowadzane są zwykle za pomocą metody wektorów wspierających (SVM) lub metod z rodziny *k* najbliższych sąsiadów [Gärtner 2003].

5. Przykład obliczeniowy

W celu zilustrowania zaprezentowanych zagadnień skonstruowano model służący do klasyfikacji opinii widzów o filmach wyświetlanych w kinach krakowskich. Analizowane teksty zaczerpnięto z serwisu http://www.kino.krakow.pl. Zamieszczane w serwisie teksty są bardzo różnorodne, często zawierają określenia slangowe, spektrum poruszanych tematów jest bardzo szerokie, teksty różnią się długością. Za przykład zróżnicowania poszczególnych dokumentów mogą służyć następujące opinie:

- *Film żenująco nudny, akcja to brak akcji, główny a zarazem jedyny wątek jest niespójny i nieciekawy, zakończenie do bani. Generalnie jedno wielkie DNO, strata czasu i pieniędzy.*
- *Nic po nim nie zostaje w głowie, najwyżej żal tych 20 złotych za bilet, totalna kaszana, szkoda mi klawiatury żeby cos więcej pisać, zastanawiam się czy nie poprosić kina o zwrot pieniędzy za bilety.*
- *Ja już oglądałam ten film i jest rewelacyjny. Brawo dla aktorów i reżysera za pomysłowość stworzenia filmu w ten naprawdę oryginalny i ciekawy sposób. Tytuł świetnie dobrany do tematyki filmu. POLECAM!!*
- *Film nawet był fajny ale chwilami mnie usypiał.*

W trakcie eksperymentu obliczeniowego wykorzystano zbiór złożony z 500 opinii. Zostały one manualnie zaklasyfikowane do trzech klas: opinie pozy-

tywne, opinie neutralne oraz opinie negatywne. Tak uzyskany zbiór potraktowano jako zbiór uczący niezbędny do skonstruowania modelu klasyfikacyjnego. Do reprezentacji informacji zawartych w przetwarzanym zbiorze dokumentów zastosowano macierz częstości. Do obliczeń zastosowano drzewa decyzyjne (obliczenia przeprowadzono za pomocą programu *Statistica*). Do ich konstruowania i oceny wykorzystano mechanizm walidacji krzyżowej. W trakcie oceny modelu okazało się, że tylko 64,8% opinii zostało zaklasyfikowanych przez model do właściwych klas (dane te dotyczą zbiorów testowych wykorzystywanych w kolejnych przebiegach walidacji krzyżowej). Może to wynikać z dużego zróżnicowania analizowanych tekstów. W dalszych badaniach zmodyfikowano strukturę modelu. W miejsce pojedynczego modelu służącego do przypisywania dokumentów do jednej z trzech klas zastosowano trzy modele, z których każdy miał na celu identyfikację dokumentów należących do jednej z rozpatrywanych klas. Stwierdzono, że takie postępowanie zwiększyło poprawność klasyfikacji dla każdej z grup (od kilkunastu do ponad dwudziestu procent). Ale uzyskiwane wyniki były w dużym stopniu uzależnione od sposobu wstępnego przetworzenia dokumentów (głównie zastosowania listy wyrazów pominiętych w trakcie analizy (stop-listy) lub też zdefiniowania listy wyrazów uwzględnionych w badaniach).

6. Wnioski

Zarówno studia literaturowe, jak i wyniki przeprowadzanych eksperymentów obliczeniowych poświęcone zagadnieniom text miningu nie pozwalają w chwili obecnej na formułowanie jednoznacznych wniosków i projektowanie procedur badawczych, lecz wskazują raczej na ogrom nie rozwiązanych jeszcze problemów. Stosunkowo dobrze rozwinęły się metody przetwarzania bazujące na reprezentacji informacji w postaci macierzy częstości. Tego typu podejście jest już wykorzystywane w praktyce dzięki modułom programowym udostępnianym przez wiodących twórców oprogramowania statystycznego. Pozostałe metody reprezentacji i przystosowane do nich metody analizy nadal pozostają raczej w sferze badań niż rzeczywistych zastosowań. Ważność prac badawczych wynika z potrzeby automatyzacji przetwarzania powszechnie występujących informacji tekstowych. Natomiast ich szeroki zakres spowodowany jest koniecznością ich niezależnego przeprowadzenia dla każdego języka naturalnego.

Literatura

Berners-Lee T. (1998), *Semantic Web Road Map*, 1998, http://www.w3.org/DesignIssues/Semantic.html.

Gärtner T. (2003), *A Survey of Kernels for Structured Data*, ACM SIGKDD Explorations Newsletter Archive, vol. 5.

Gruber T.R. (1993), *A Translation Approach to Portable Ontology Specifications*, http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.

Hearst M.A. (1999), *Untangling Text Data Mining*, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html.

Montes-y-Gómez M. i in. (2000), *Comparison of Conceptual Graphs*. Proceedings MICAI-2000, 1st Mexican International Conference on Artificial Intelligence, Acapulco, [w:] O. Cairo, L.E. Sucar, F.J. Cantu (red.), *MICAI 2000: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence N 1793, Springer, s. 548-556.

Nierman A. (2002), *Evaluating Structural Similarity in XML Documents*, Proceedings of the Fifth International Workshop on the Web and Databases, WebDB.

Oldakowski R., Bizer C. (2005), *SemMF: A Framework for Calculating Semantic Similarity of Objects Represented as RDF Graphs*, Freie Universität Berlin, Institut für Produktion, Berlin.

APPLICATION OF TEXTUAL DATA IN MODELLING AND CLASSIFICATION PROBLEMS

Summary

The main subject of this paper is related to text mining analysis. It presents a problem of information retrieval from text documents, information representation, and choice of correct methods of analysis. It was shown that the choice of proper information representation method is the most important stage during research. The paper presents the present abilities of text mining and direction of its development.