TEXT MINING JAKO NARZĘDZIE POZYSKIWANIA INFORMACJI Z DOKUMENTÓW TEKSTOWYCH

Paweł Lula
Akademia Ekonomiczna w Krakowie, Katedra Informatyki

Stale i szybko rosnące zasoby informacyjne są jedną z cech współczesności. Ich prawidłowe pozyskanie, przeanalizowanie i przetworzenie jest warunkiem koniecznym do prawidłowego funkcjonowania każdego człowieka. To trudne zadanie jest w dużym stopniu wspomagane przez najnowsze zdobycze nauki i techniki, czego najlepszym przykładem jest rola, jaką odgrywa obecnie analiza danych i informatyka.

Należy jednak zauważyć, że wśród istniejących i obecnie tworzonych zasobów informacyjnych znaczną rolę odgrywają dokumenty tekstowe. Ich analiza i przetworzenie jest stosunkowo prosta dla człowieka, jednak próby automatyzacji tych zadań są szczególnie trudne. Podstawową przyczyną trudności jest niski stopień ustrukturyzowania tekstów i brak jednoznacznych metod ich interpretacji.

Przeprowadzane w wielu ośrodkach badania wskazują jednak, że zarezerwowana przez długi okres czasu wyłącznie dla człowieka działalność związana z przetwarzaniem dokumentów tekstowych może być w dużym stopniu zautomatyzowana. Systemy komputerowe wyposażone w odpowiednie oprogramowanie są w stanie gromadzić olbrzymie zasoby tekstowe, pozyskiwać zawarte w nich informacje i dokonywać sprawnego ich przetworzenia.

Podstawowym celem niniejszego opracowania jest prezentacja eksploracyjnej analizy dokumentów tekstowych (określanej jako *text mining*), która rozumiana jest jako zbiór koncepcji, metod oraz zaimplementowanych w postaci programów komputerowych algorytmów przetwarzania zasobów tekstowych, prowadzących do zautomatyzowania procesów przetwarzania dokumentów sporządzonych w językach naturalnych. W kolejnych częściach tekstu ukazane zostaną:

- definicja i cele zastosowań text miningu oraz powiązania z dziedzinami pokrewnymi,
- typy problemów poznawczych rozpatrywanych na gruncie text miningu wraz z krótkim omówieniem przykładowych zastosowań,
- metody reprezentacji informacji zawartych w dokumentach tekstowych,
- podstawowe metody przetwarzania informacji pozyskanych z dokumentów tekstowych.

Ilustracją dla omawianych zagadnień będzie prezentacja analizy przykładowego zestawu dokumentów. Wszystkie obliczenia zostaną wykonane za pomocą modułu *Text Miner*, będącego częścią składową pakietu *STATISTICA*.

Czym jest text mining?

Pojęcie text miningu zdobywa popularność od końca lat dziewięćdziesiątych XX wieku. Do jego twórców należy zaliczyć Marti A. Hearst, która w swoim powszechnie cytowanym artykule ([10]) definiuje *text mining* jako *proces mający na celu wydobycie z zasobów tekstowych nieznanych wcześniej informacji*.

Text mining ma charakter interdyscyplinarny, gdyż korzysta ze zdobyczy badaczy zajmujących się *data miningiem*, uczeniem maszynowym, przetwarzaniem języka naturalnego (a w szczególności osób zajmujących się metodami wyszukiwania i pozyskiwania informacji oraz tłumaczenia maszynowego), statystyką, lingwistyką oraz informatyką. W tym miejscu warto poddać krótkiej dyskusji podobieństwa i różnice pomiędzy text miningiem a każdą z wyróżnionych powyżej dziedzin.

Nietrudno zauważyć związki text miningu z data miningiem. W obu przypadkach analityk dąży do odkrycia nieznanych wcześniej prawidłowości poprzez eksplorację dużych zasobów danych. Jednakże charakter przetwarzanych zasobów jest inny: podejście data miningowe przystosowane jest do analizy danych o ściśle określonej strukturze, o wartościach wyrażonych na klasycznych skalach pomiarowych; natomiast analiza text miningowa ukierunkowana jest na analizę tekstu, a więc zasobu nie mającego określonej struktury lub o strukturze określonej w sposób nieprecyzyjny i dość dowolny (jak na przykład w plikach HTML). Jednakże text mining w dużym stopniu korzysta z *data miningu* - zapożyczył eksploracyjne podejście do procesu analizy, ukierunkowanie procesu analizy na zastosowania, przywiązywanie dużej wagi do prostoty interpretacji i użyteczności uzyskanych wyników oraz wykorzystanie wspólnego (w dużej części) zestawu metod i narzędzi.

Można wskazać na wiele punktów wspólnych pomiędzy text miningiem a *uczeniem maszynowym*, którego istnienie (podobnie jak text miningu) byłoby niemożliwe bez stosowania metod i narzędzi informatyki. Głównym założeniem uczenia maszynowego jest wykrycie istniejących prawidłowości poprzez analizę wielu przykładów jej realizacji, a następnie ujęcie wykrytych w ten sposób reguł w postaci modelu łatwego do wykorzystania przez system komputerowy. Do najczęściej stosowanych typów modeli należy zaliczyć sieci neuronowe, drzewa decyzyjne czy algorytmy genetyczne. Uczenie maszynowe jest często stosowane w text miningu. Jednakże nie zawsze, gdyż w wielu przypadkach stosowane jest podejście odmienne niż w uczeniu maszynowym, polegające na próbie dokonania ekstrakcji informacji za pomocą klasycznych metod statystycznych czy też poprzez zastosowanie wyrafinowanych modeli lingwistycznych.

Text mining korzysta również z olbrzymich osiągnięć badaczy zajmujących się *przetwarzaniem języka naturalnego* (*NLP - Natural Language Processing*). Za pioniera w tej

dziedzinie uznaje się wybitnego lingwistę Noama Avrama Chomsky'ego (ur. 1928), który był przekonany o istnieniu ogólnych reguł gramatycznych pozwalających na stworzenie takiego matematycznego (formalnego) modelu, który umożliwi rozumienie i tworzenie poprawnych zdań w języku naturalnym. W rozumieniu Chomsky'ego gramatyka to zbiór symboli i zbiór reguł określających sposób ich przetwarzania. Mimo ogromnych osiągnięć lingwistyki formalnej (widocznych np. w technikach kompilacji wykorzystywanych przy programowaniu komputerów) obecny poziom wiedzy wskazuje, że bogactwo języków naturalnych nadal wymyka się ograniczeniom nakładanym przez ich formalny opis. Prace Chomsky'ego kontynuowane były przez wielu jego następców i obejmowały coraz szerszy zakres tematyczny.

W kręgu zainteresowań badaczy znalazło się między innymi zagadnienie wykorzystania języka naturalnego jako *środka komunikacji* pomiędzy człowiekiem a maszyną. Spełnienie tego postulatu wymagało, aby komputer posiadał zdolność rozumienia ludzkiego języka oraz aby potrafił samodzielnie generować wypowiedzi, stosując ten sam sposób komunikacji. Znaczenie, jakie przywiązywano do posiadania przez komputery umiejętności posługiwania się językiem naturalnym było tak duże, że Alan Turing uznał ją za kryterium wystarczające do przypisania maszynie cech charakterystycznych dla istot inteligentnych. Rozpatrując zagadnienie komunikacji z maszyną przy wykorzystaniu języka naturalnego, należy wspomnieć o osiągnięciach Josepha Weizenbauma, twórcy programu ELIZA, posiadającego umiejętność prowadzenia rozmowy (za pośrednictwem klawiatury i monitora) na pozornie dowolny temat. Program ELIZA stał się pierwowzorem dla wielu dalszych prac zmierzających do utworzenia jeszcze doskonalszych narzędzi rozumiejących, a przynajmniej inteligentnie przetwarzających mowę ludzką. Program ELIZA, jak i wszystkie kolejne następne osiągnięcia w tej dziedzinie, nie osiągnęły zdolności posługiwania się językiem naturalnym na poziomie człowieka. Nie spełniły więc podstawowego kryterium wymienianego przez Turinga jako warunku koniecznego do przypisania maszynie cech istot inteligentnych. Można wskazać na istotne podobieństwa istniejące pomiędzy celem realizacji analizy text miningowej a przedstawionymi powyżej kierunkami badań nad przetwarzaniem języka naturalnego. W obu przypadkach system komputerowy musi być w stanie zrozumieć przekaz tekstowy. Od typowego systemu NLP oczekuje się, że będzie on w stanie prawidłowo zareagować na każdą pojedynczą wypowiedź, system text miningowy ukierunkowany jest raczej na analizę i przetworzenie dużych zasobów informacyjnych. W tym drugim przypadku nie oczekuje się również generowania rezultatów w postaci przekazu w języku naturalnym (często wręcz odwrotnie - pożądane jest pojawienie się na wyjściu informacji w postaci ustrukturyzowanej).

Innym istotnym kierunkiem badań dotyczących przetwarzania języka naturalnego jest *tłumaczenie maszynowe* (*MT - Machine Translation*), mające na celu utworzenie systemu komputerowego tłumaczącego bez ingerencji człowieka pełne zdania pomiędzy różnymi językami. Początkowo tego typu systemy budowano w oparciu o reguły o charakterze lingwistycznym, określające sposób dokonywania translacji. Mimo osiągniętych sukcesów jakość tak działających systemów nie zawsze była wystarczająca. W wielu przypadkach lepsze rezultaty można uzyskać, stosując podejście alternatywne, określane jako tłumaczenie maszynowe oparte na przykładach (*Example-based machine translation*),

zaproponowane na początku lat osiemdziesiątych XX wieku ([15]), ale rozwijane dopiero od lat dziewięćdziesiątych poprzedniego stulecia. Zaproponowane rozwiązanie polega na stworzeniu dużej bazy par zdań o identycznym znaczeniu, ale zapisanych w dwóch różnych językach. Następnie, wykorzystując ideę uczenia maszynowego - system komputerowy, analizując odpowiadające sobie zdania, uczy się reguł translacji. W trakcie funkcjonowania systemu w istniejącej bazie wyszukiwane jest zdanie najbardziej zbliżone do tego, które ma zostać przetłumaczone, a następnie - wykorzystując znajdujący się w bazie odpowiednik - generowane jest zdanie w języku docelowym. Wykorzystywany w tego typu rozwiązaniach mechanizm uczenia maszynowego jest wspomagany przez dodatkową, wprowadzoną przez człowieka wiedzę. Może ona mieć charakter lingwistyczny ([18]) lub statystyczny ([2]). Poszukując sposobów polepszenia jakości działania opracowywanych rozwiązań, zaproponowane zostały podejścia w dużym stopniu zbliżone do technik data i text miningowych. Pierwsze z nich polegało na łącznym stosowaniu wielu niezależnych mechanizmów tłumaczących i generowaniu odpowiedzi ostatecznej jako wypadkowej proponowanych rozwiązań cząstkowych. Inne bazowało na opracowanym wcześniej formalnym modelu (ontologii) dziedziny związanej z tematyką tekstu i dążyło do stworzenia reprezentacji przeznaczonej do tłumaczenia wypowiedzi za pomocą pojęć dostępnych w przyjętej ontologii, a następnie wygenerowanie odpowiedników reprezentowanych faktów w języku docelowym. Przegląd metod tłumaczenia maszynowego, próbę ich oceny i perspektywy rozwoju zostały przedstawione w pracy: [19].

W spektrum badań nad przetwarzaniem języka naturalnego znalazły się również zagadnienia dotyczące *ekstrakcji informacji* z dokumentów tekstowych. Prace w tym zakresie prowadzone są od przełomu lat osiemdziesiątych i dziewięćdziesiątych XX wieku. W latach 1987-1998 prace sponsorowane były przez agencję DARPA (*Defence Advanced Research Projects Agency*) w ramach konferencji MUC (*Message-Understanding Conference*). Celem było opracowanie metod pozyskiwania ściśle określonych informacji i wstawienie ich do wcześniej zaprojektowanej struktury ramowej (szablonu). Opracowane podejście służyło do analizy przesyłanych komunikatów wojskowych, wiadomości agencyjnych, ekonomicznych oraz wykrywania przejawów działalności terrorystycznej. W proponowanych rozwiązaniach szczególną uwagę zwracano na mechanizmy pozwalające na identyfikację nazw własnych oraz wykrywanie koreferencji (czyli różnych określeń odnoszących się do tego samego obiektu). Wypracowane w tym zakresie metody należą obecnie do grupy najpopularniejszych algorytmów text miningowych, a obszar ich zastosowań ulega ciągłemu rozszerzeniu.

Dość często text mining utożsamiany jest z zaawansowanymi metodami *wyszukiwania informacji* (stosowanymi na przykład w internetowych serwisach wyszukujących). Według części badaczy pogląd taki nie jest jednak uzasadniony, gdyż obie wymienione formy analizy mają inne cele. Celem metod i narzędzi wyszukujących jest zwrócenie ***zestawu dokumentów*** spełniających podane kryteria, zaś text mining ma dostarczać ***informacje*** pozyskane w wyniku analizy dokumentów, a nie same dokumenty. W tym kontekście za bardzo ważny problem należy uznać sposób reprezentacji informacji pozyskanych w wyniku analizy. Wpływa on w istotny sposób na dalszy przebieg analizy.

Text mining, jak każda dziedzina analizy danych, korzysta z wielowiekowych zdobyczy *statystyki*, która w odniesieniu do otaczającej rzeczywistości powinna opisywać, wyjaśniać prawidłowości i prognozować. Na tym tle text mining należy traktować jako szczególny obszar zastosowań statystyki, w wielu miejscach nietypowy i trudny, ale przez to szczególnie ciekawy i obiecujący. Szczególną popularnością wśród badaczy zjawisk natury językowej cieszą się metody statystyki opisowej, łańcuchy Markowa i modele probabilistyczne.

Związki text miningu z *informatyką* mają charakter wieloaspektowy. Omawiając ten problem, należało by wspomnieć o teorii algorytmów i struktur danych, o sztucznej inteligencji czy wreszcie o bazach danych. Zaangażowanie komputerów do przetwarzania danych tekstowych ma stosunkowo długą historię, rozpoczynającą się w latach pięćdziesiątych poprzedniego stulecia. Algorytmy służące ich przetwarzaniu były tematem rozważań największych autorytetów w zakresie algorytmiki (N. Wirth, D. Knuth, van Tassel). Współczesny text mining wykorzystuje w dużym stopniu zdobycze informatyki w zakresie projektowania, analizy i przetwarzania struktur danych. Pozwalają one rozwiązać problemy związane ze sposobem przechowywania pozyskanych informacji w systemach komputerowych. Wiele problemów text miningu jest dobrze znanych specjalistom z zakresu baz danych. Na styku tych dziedzin rozważane są problemy przechowywania dużej ilości tekstów, realizacji operacji indeksowania czy też definiowania zapytań. Bardzo mocnym ogniwem łączącym text mining z informatyką są również omówione już wcześniej zagadnienia przetwarzania języka naturalnego, które z punktu widzenia informatyki postrzegane są jako wiodący kierunek badań w ramach sztucznej inteligencji.

Zastosowania text miningu

Zakres zastosowań text miningu jest bardzo szeroki. Próbę określenia najważniejszych obszarów zastosowań można znaleźć w wielu pozycjach literaturowych (np.: [6]). Zestawienie zaprezentowane w dalszej części bieżącego punktu nie ma z całą pewnością charakteru całościowego, jest jedynie przeglądem najpopularniejszych zastosowań.

Pozyskiwanie informacji z dokumentów

Mechanizmy pozyskiwania informacji bazują przede wszystkim na próbie dopasowania do poszczególnych fragmentów tekstu wzorców określających rodzaj poszukiwanych treści. Najprostszym typem wzorców może być zbiór słów kluczowych. Możliwe jest tworzenie wzorców służących identyfikacji fraz określających typ zdarzenia, czas, lokalizację, wykonawcę, skutki itd. Pobrane w ten sposób informacje, po nadaniu im ściśle określonej struktury, mogą być umieszczane w klasycznych bazach danych, dzięki czemu mogą być w dogodny sposób poddawane dalszemu przetworzeniu.

W ostatnim okresie można obserwować dynamiczny rozwój tego typu zastosowań w odniesieniu do analizy doniesień literaturowych z zakresu biologii. W tej dziedzinie wiedzy w przeciągu ostatnich kilkudziesięciu lat daje się zauważyć wykładniczy wzrost liczby publikacji naukowych ([17]). Fakt ten powoduje, że zapoznanie się badaczy z bieżącą literaturą dotyczącą ich sfery zainteresowań jest niemożliwe. Dlatego z dużą

nadzieją przyjmowane są rozwiązania pozwalające na automatyzację tego czasochłonnego procesu. Podstawowym celem realizowanych prac jest stworzenie systemu wspomagającego proces pozyskiwania informacji z literatury przedmiotu, nadanie im zdefiniowanej struktury i umieszczanie ich w bazie danych ([13], [24]), co znacznie ułatwi i przyspieszy ich dalsze przetworzenie.

Innym, nie mniej ważnym, obszarem zastosowań metod pozyskiwania informacji z dokumentów tekstowych są systemy wspierające działalność biznesową. W charakterze przykładu można wskazać na możliwości ich zastosowań w systemach CRM, w których mogą służyć jako mocne narzędzie analizy danych tekstowych dotyczących klientów firmy (poznanie profilu klienta, prognozowanie przejścia klienta do konkurencji, identyfikacja przyczyn odejścia klienta [20]. Tego typu zastosowania mogą być również pomocne w bankowości, gdzie mogą służyć do analizy korespondencji z klientami banku w celu określenia prawdopodobieństwa niespłacenia zaciągniętego kredytu [2].

Identyfikacja wiadomości zawierających określone treści

W tym przypadku celem analizy jest stworzenie systemu monitorującego dużą liczbę dokumentów w celu identyfikacji tych, które mogą być istotne z punktu widzenia zdefiniowanego kryterium. W przypadku opisywanego zastosowania zakłada się, że automatyzacji podlega jedynie wskazanie potencjalnych dokumentów, nie zaś ich analiza, która powinna być dokonana przez człowieka.

W pracy [21] omówione są możliwości zastosowań tego typu systemu w systemach informacyjnych przedsiębiorstw w charakterze podstawowego narzędzia stosowanego przy przetwarzaniu informacji przechowywanych w tekstowych hurtowniach danych. Z uwagi na duży udział informacji tekstowych w ogólnych zasobach informacyjnych wykorzystywanych w działalności biznesowej (szacowany na około 80%) budowa hurtowni przechowujących dokumenty tekstowe staje się koniecznością. Tego typu hurtownia powinna gromadzić między innymi: raporty, tworzone dokumenty, notatki, korespondencję, doniesienia prasowe. Metody text miningu umożliwiają wyszukiwanie informacji, pozyskiwanie dokumentów według zdefiniowanych kryteriów, określenie charakteru dokumentu (tematyka, kluczowe zagadnienia, język sporządzenia), organizowanie struktury dokumentów, wizualizację zgromadzonych zasobów, konstruowanie profilu klienta (poprzez analizę przesyłanych przez niego zapytań, próśb i uwag).

Wypracowane na tym polu rozwiązania cieszą się dużą popularnością wśród milionów użytkowników Internetu, gdyż pozwalają na identyfikację i automatyczne usuwanie spamu. Są one również stosowane do wykrywania przejawów działalności niezgodnej z prawem i stanowiącej zagrożenie dla bezpieczeństwa publicznego.

Generowanie streszczeń

Widoczna prawie w każdej dziedzinie życia gwałtownie zwiększająca się liczba dokumentów powoduje, że zapoznanie się z nimi w sposób całościowy jest praktycznie niemożliwe. Częściowym rozwiązaniem tego problemu jest analiza streszczeń dokumentów. Próby

zautomatyzowania systemu generowania streszczeń należą do jednych z głównych zastosowań rozwiązań text miningowych. Jest to zadanie trudne, gdyż jego realizacja wymaga pozyskania z dokumentu najistotniejszych faktów (często o nieznanej wcześniej strukturze) i ich wyrażeniu za pomocą poprawnie sformułowanych zdań w języku naturalnym. Najprostsze rozwiązania tego typu polegają na identyfikacji za pomocą metod statystycznych najistotniejszych słów kluczowych, ale takie rozwiązanie nie pozwala na poinformowanie o charakterze istniejących związków pomiędzy poszczególnymi wyrazami. Inne podejście zakłada, że streszczenie składać się będzie z najważniejszych zdań pobranych bezpośrednio z tekstu źródłowego. W tym przypadku zachodzi potrzeba dokonania wyboru zdań najlepiej reprezentujących poddawany analizie dokument. Podejmowane są różnorodne próby określenia ważności zdań. Pierwsza grupa rozwiązań polega na agregacji wskaźników ważności dla poszczególnych wyrazów wchodzących w skład zdania i przyjęciu uzyskanej wartości jako oceny istotności zdania. Następnie zdania porządkowane są według malejącej wartości informacyjnej. Te najistotniejsze umieszczane są w streszczeniu. Inne podejście zakłada, że w streszczeniu umieszczane będą te zdania, które zawierają pewne zwroty wskazujące na ich dużą wartość informacyjną (za przykład posłużyć mogą: „największe osiągnięcie", „podsumowując", „istotne rezultaty").

Klasyfikacja wzorcowa

Zadanie klasyfikacji wzorcowej polega na analizie zbioru dokumentów i przypisaniu każdego z nich - biorąc pod uwagę zawarte w nich informacje - do jednej z wcześniej wyróżnionych klas. Realizacja tego zadania wymaga zdefiniowania wzorców poszczególnych klas i określenia sposobu pomiaru podobieństwa analizowanego dokumentu do poszczególnych wzorców w celu ustalenia najlepszego dopasowania. Zasadniczym czynnikiem wpływającym na sposób realizacji tych zadań jest przyjęty sposób reprezentacji informacji zawartych w dokumencie.

Klasyfikacja bezwzorcowa (grupowanie, klasteryzacja)

W przypadku analizy dokumentów tekstowych realizacja zadania klasyfikacji bezwzorcowej sprowadza się do określenia jednorodności zestawu dokumentów, wydzieleniu grup dokumentów podobnych, określeniu zależności pomiędzy grupami oraz scharakteryzowaniu cech charakterystycznych dla dokumentów wchodzących w skład wydzielonych skupień. Podobnie jak w przypadku klasyfikacji wzorcowej, również przy rozwiązywaniu zagadnień grupowania najistotniejszym problemem jest określenie liczbowej miary podobieństwa pomiędzy dokumentami. Jej zdefiniowanie jest warunkiem koniecznym do realizacji zadania grupowania przy wykorzystaniu wysoko ocenianych klasycznych metod klasyfikacji (np.: metod taksonomicznych, klasyfikujących sieci neuronowych).

Identyfikacja powiązań

Identyfikacja powiązań może być rozumiana jako wykrycie związków istniejących pomiędzy informacjami pozyskanymi z dokumentów tekstowych lub też jako identyfikacja dokumentów, które są powiązane ze sobą ze względu na zawarte w nich treści.

W przypadku pierwszego typu problemu możliwe jest wyróżnienie dwóch zasadniczych etapów realizowanego procesu. Pierwszym z nich jest pozyskanie informacji z dokumentu źródłowego, a drugim identyfikacja nieznanych wcześniej powiązań istniejących pomiędzy odkrytymi faktami. W drugim przypadku chodzi wyłącznie o wskazanie zbioru dokumentów powiązanych z rozpatrywanym dokumentem lub wskazanym tematem. Zakłada się, że dalsza analiza zostanie wykonana przez człowieka.

W literaturze prezentowane są różne zastosowania systemów identyfikacji powiązań. Za przykład może posłużyć system zautomatyzowanej analizy raportów likwidatorów szkód w celu wykrycia przyczyn zwiększania się zgłaszanych przez klientów szkód ([5]). Informacje pozyskane z tekstów, wykorzystane łącznie z danymi ilościowymi, pozwoliły na rozpoznanie słabych punktów istniejących rozwiązań i wykrycie źródeł nadużyć.

Wizualizacja

Zadanie wizualizacji jest zwykle powiązane z próbą rozwiązania innego typu zadania. Jego głównym celem jest zapewnienie użytkownikowi prostej metody interpretacji uzyskanych wyników. Najczęściej wizualizacji poddawane są związki zachodzące pomiędzy wyodrębnionymi faktami lub zależności zachodzące w strukturze rozpatrywanego zbioru dokumentów.

Generowanie odpowiedzi na pytania

Tego typu zagadnienia, będące głównie w sferze zainteresowań sztucznej inteligencji, dotyczą badań nad możliwością zrozumienia przez maszynę pytania zadanego przez człowieka i sformułowanego w języku naturalnym. Prowadzone na tym polu prace mają na celu określenie możliwości wykorzystania języka naturalnego jako narzędzia budowy interfejsu do zaawansowanych systemów informacyjnych.

Przebieg analizy text miningowej i charakterystyka stosowanych metod

Proces text miningowej analizy dokumentów tekstowych jest wieloetapowy. Do jego głównych etapów należy zaliczyć:

- określenie celu, zakresu i kosztów analizy,
- przekształcenie dokumentów źródłowych do postaci dogodnej do dalszego przetworzenia,
- przeprowadzenie obliczeń mających na celu zrealizowanie ustalonego celu analizy,
- interpretacja wyników.

Określenie celu, zakresu i kosztów analizy

Określenie celu badań jest jednym z najważniejszych etapów procesu analizy. Prawidłowe zrealizowanie tego etapu warunkuje cały dalszy przebieg prac. Cel wpływa na postać

formułowanych hipotez, zbiór stosowanych metod, na sposób przygotowania danych czy też na sposób prezentacji wyników. Zakres jest wypadkową celu, wybranych metod i dostępnego budżetu. Koszty mogą mieć różnorodny charakter. Ich dokładna analiza stanowi podstawę do podjęcia decyzji o realizacji badań, czy też rozszerza lub ogranicza ich zakres. Szczególnie istotne jest porównanie kosztów analizy z wynikającymi z jej przeprowadzenia korzyściami.

Przekształcenie dokumentów źródłowych do postaci dogodnej do dalszego przetworzenia

W ramach realizacji opisywanego w tym miejscu punktu analizy są dwie czynności: **wstępna obróbka analizowanego zbioru dokumentów** oraz **określenie sposobu reprezentacji informacji występujących w dokumencie**.

Wstępna obróbka analizowanego zbioru dokumentów

Wstępna obróbka dokumentów obejmuje między innymi:

- transformację dokumentów do postaci tekstowej (znaczna liczba programów pozwalających na analizę tekstów operuje wyłącznie na plikach tekstowych),
- 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 naszego języka nie jest ujednolicony).

W wyniku realizacji powyższego etapu analizy uzyskiwany jest zestaw plików tekstowych, zawierających teksty bez jakichkolwiek informacji formatujących, o ujednoliconym sposobie kodowania znaków.

Przebieg całego dalszego procesu analizy uzależniony jest od podejmowanej w tym momencie decyzji dotyczącej sposobu reprezentacji informacji zawartych w tekście.

Określenie sposobu reprezentacji informacji występujących w dokumencie

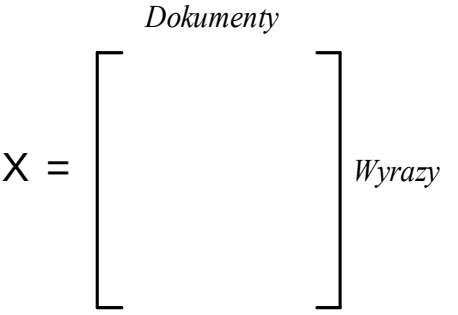
Ten bardzo ważny etap analizy text miningowej ma na celu określenie sposobu dalszego przechowywania w systemie komputerowym informacji pochodzących z przetwarzanej kolekcji dokumentów. Istnieją dwa podstawowe podejścia do rozwiązania tego zagadnienia. Pierwsze z nich zakłada, że w systemie komputerowym przechowywane są wyłącznie informacje o częstości występowania poszczególnych wyrazów w przetwarzanych dokumentach. Głównym założeniem drugiego rozwiązania jest dążenie do nadania informacjom pozyskanym z dokumentów postaci ustrukturyzowanej. Jedno i drugie rozwiązanie ma swoje mocne i słabe strony, które należy dokładnie poznać, aby w sposób świadomy dokonać właściwego wyboru sposobu reprezentacji pozyskanych z dokumentów informacji.

Reprezentacja bazująca na liście słów występujących w dokumencie

Metoda ta zakłada, że strukturą reprezentującą dokument jest wektor, którego poszczególne elementy informują o liczbie wystąpienia poszczególnych słów. Prace zmierzające do uzyskania tego typu reprezentacji rozpoczynają się od usunięcia z każdego przetwarzanego dokumentu znaków interpunkcyjnych i utworzenia, niezależnie dla każdego przetwarzanego tekstu, listy występujących w nim słów. Listy wyrazów związane z poszczególnymi dokumentami poddawane są 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 określana jest 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 przy wykorzystaniu słowników.

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



Rys. 1. Struktura macierzy częstości

Wierszom macierzy odpowiadają kolejne słowa występujące w dokumentach, zaś kolumny 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 wyrażane jest poprzez określenie podobieństwa odpowiadających im wierszy macierzy **X**, natomiast 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 posiadają wartości niezerowe na więcej niż jednej pozycji. Wiersze odpowiadające wyrazom występującym tylko w jednym dokumencie są usuwane z macierzy.

Wyznaczone częstości wystąpień poddać można transformacji służącej uwypukleniu cech wspólnych dokumentów. Potrzebę przeprowadzenia transformacji uzasadniać może spostrzeżenie mówiące, że o stopniu podobieństwa pomiędzy dokumentami mocniej świadczy sam fakt wystąpienia takich samych słów niż precyzyjnie określona ilość ich wystąpień. Do najczęściej stosowanych metod transformacji macierzy częstości należy:

- *reprezentacja binarna* - zakłada, że rejestrowany jest sam fakt wystąpienia *i*-tego słowa w *j*-tym dokumencie, natomiast nie precyzuje się 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.
- *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 sam fakt wystąpienia słowa w dokumencie i tylko w niewielkim stopniu uwzględnia liczbę wystąpień tego wyrazu.
- *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, natomiast z punktu widzenia możliwości klasyfikacji szczególnie przydatne są wyrazy występujące w stosunkowo niewielkiej liczbie dokumentów. Podejście takie znajduje swoje odzwierciedlenie w stosowaniu formuły ([12]): $\left(\left(1+\log\left(x_{ij}\right)\right)*\log\left(\frac{N}{df_i}\right)\right)$ dla elementów niezerowych macierzy częstości i pozostawienie elementów zerowych na dotychczasowym poziomie (w przedstawionej formule wartość N określa liczbę wszystkich dokumentów, zaś df_i jest liczbą dokumentów zawierających *i*-ty wyraz).

Reprezentacja w postaci ustrukturyzowanej

Podstawową zaletą przedstawionego powyżej sposobu konstrukcji numerycznej reprezentacji tekstów jest prostota realizacji niezbędnych obliczeń oraz duży wybór metod odpowiednich do dalszego przetwarzania tak uzyskanej struktury. Natomiast wadą jest nieuwzględnienie wielu dostępnych w tekście źródłowym informacji. Najważniejszą przyczyną strat jest uwzględnienie wyłącznie informacji o występujących wyrazach, bez analizowania tworzonych przez nie struktur językowych.

Całkowicie odmienne są cechy charakteryzujące reprezentację informacji pozyskanych z dokumentu w postaci ustrukturyzowanej. Podstawowym założeniem jest chęć przeciwdziałania utracie informacji właściwej dla omówionego powyżej rozwiązania poprzez zastosowanie struktur danych odpowiednich do przechowywania: związków wynikających z kolejności wyrazów, charakterystyk opisywanych w dokumentach obiektów, relacji

pomiędzy nimi oraz zależności przyczynowo-skutkowych. Strukturami dogodnymi do realizacji tych postulatów mogą być łańcuchy znaków, listy, rekordy, drzewa i grafy. Niestety stosowanie tego typu rozwiązań napotyka na szereg problemów. Pierwszym z nich są trudności związane z przekształceniem dokumentu tekstowego do postaci ustrukturyzowanej. Zaś drugim jest znacznie mniejszy wachlarz metod analitycznych przystosowanych do operowania na informacjach przechowywanych przy wykorzystaniu złożonych struktur danych.

Pierwsze prace zmierzające do nadania struktury informacjom wyodrębnionym z dokumentów zmierzały do umieszczenia ich w rekordach tabel wchodzących w skład relacyjnych baz danych. Realizacja tego zamierzenia pozwala na dalszą eksplorację informacji za pomocą dobrze znanych i bardzo wydajnych mechanizmów zapytań, bazujących przede wszystkim na języku SQL.

Inne podejście postulowało przejście od struktur relacyjnych do drzewiastych i przechowywanie informacji w postaci dokumentów zdefiniowanych w języku XML. Podejście to wykorzystywało główną cechę języka XML, jaką jest możliwość definiowania własnych znaczników i precyzyjnego określania relacji pomiędzy poszczególnymi znacznikami.

Dalszym rozwinięciem są sposoby reprezentacji bazujące na modelu rozpatrywanego w dokumencie fragmentu rzeczywistości. Chęć wykorzystania tego typu podejścia wymaga przyjęcia założenia o identycznym zakresie dziedzinowym wszystkich przetwarzanych jednocześnie tekstów oraz o istnieniu, dla rozpatrywanej dziedziny zainteresowań, formalnych narzędzi opisu rzeczywistości. W tej dziedzinie badań obecnie największą uwagę poświęca się tematyce ontologii, rozumianej jako formalny sposób opisu wyodrębnionego fragmentu rzeczywistości. Za przykład pionierskiej pracy w tym zakresie posłużyć może [8]. Definicja ontologii obejmuje opis obiektów występujących w rzeczywistości oraz opis zależności pomiędzy nimi. Obiekty reprezentowane są przez struktury danych zwane encjami. Mają one zwykle strukturę hierarchiczną, co powoduje, że reprezentowane są w postaci struktur drzewiastych. Wśród zależności występujących pomiędzy reprezentowanymi w ontologii obiektami szczególnie ważną rolę odgrywają relacje semantyczne (typu: *posiada*, *jest częścią składową* itd.). Z tego powodu układ obiektów wraz z opisem występujących pomiędzy nimi zależności semantycznych nazywany jest *siecią semantyczną*. Strukturami dogodnymi do reprezentacji sieci semantycznej są grafy.

Niestety w ogólnym przypadku przekształcenie informacji zawartych w tekście do postaci ustrukturyzowanej (nawet w postaci prostej tabeli) nie jest zadaniem łatwym. Proces ten może być realizowany w sposób manualny, ale wówczas jest zadaniem bardzo czasochłonnym. Może się odbywać za pomocą reguł zdefiniowanych przez ekspertów, ale zwykle reguły obejmują tylko najbardziej typowe przypadki i w niewielkim stopniu są w stanie przetworzyć w poprawny sposób informacje w układzie nawet w niewielkim stopniu odbiegającym od standardów uwzględnionych przez eksperta. Może w końcu być realizowane za pomocą uczenia maszynowego, ale wymaga dużych zbiorów uczących.

Za szczególnie obiecujące (a jednocześnie stosunkowo trudne i wymagające) należy uznać zastosowanie koncepcji sieci semantycznych w połączeniu z text miningiem. Text mining

może zostać wykorzystany jako narzędzie wspomagające budowę ontologii, dzięki analizie statystycznej tekstów prowadzącej do identyfikacji najistotniejszych obiektów i związków (przy wykonywaniu tego typu zadań niezbędne jest korzystanie z informacji słownikowych). Drugi obszar zastosowań sprowadza się do wydobywania informacji wymaganych do konstrukcji sieci semantycznej.

Przeprowadzenie obliczeń

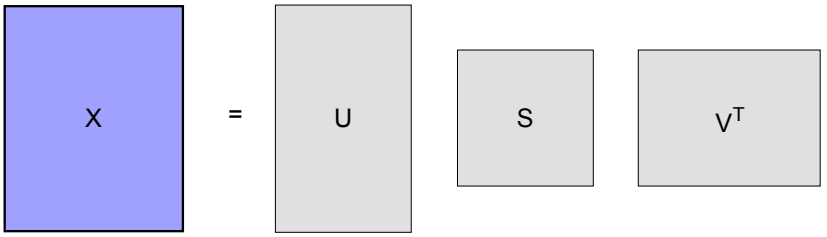
Przystępując do realizacji zasadniczego etapu analizy text miningowej, należy dokonać wyboru właściwych metod analizy. W trakcie selekcji metod należy brać pod uwagę przede wszystkim cel analizy. Należy również uwzględnić przyjęty sposób reprezentacji informacji pochodzących z zasobów tekstowych.

Jednym z najbogatszych zbiorów publikacji dotyczącym metod text miningowych (przy szczególnym uwzględnieniu zagadnienia pozyskiwania informacji) są materiały z konferencji *The Text REtrieval Conference* (TREC) dostępne pod adresem: http://trec.nist.gov/.

W dalszej części bieżącego fragmentu tekstu zostanie przedstawiona krótka charakterystyka najpopularniejszych metod text miningowych.

Metoda oparta na rozkładzie według wartości osobliwych (SVD)

Jedną z trudności pojawiających się przy 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 [4]. Opiera się ona na zastosowaniu w odniesieniu do macierzy częstości (w postaci oryginalnej lub przekształconej) rozkładu według wartości osobliwych (schemat pokazuje rys. 2).



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

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) reprezentowany jest przez jeden wiersz odpowiedniej macierzy, przy czym (podobnie jak w analizie głównych

składowych) poszczególne współrzędne uporządkowane są malejąco według ich wartości informacyjnej, co pozwala na redukcję przestrzeni poprzez uwzględnienie tylko pewnej liczby początkowych elementów wektora. Zajmując się klasyfikacją 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, gdyż pozwoliłoby to na graficzną reprezentację struktury zbioru dokumentów.

Metoda LSA jest dogodnym narzędziem przetwarzania informacji reprezentowanych w postaci listy wyrazów. Jej wyniki stanowią zwykle punkt wyjścia do dalszych obliczeń.

Taksonomiczne metody grupowania

Metody taksonomiczne należą do klasycznych metod analizy danych. Służą do grupowania obiektów charakteryzowanych za pomocą przyjętego zbioru atrybutów. W charakterze danych wejściowych wprowadzana jest zwykle macierz odległości albo macierz podobieństwa pomiędzy analizowanymi obiektami. Metody taksonomiczne mogą być stosowane w odniesieniu do analizy informacji pochodzących z dokumentów tekstowych, jeśli tylko dla przyjętego sposobu reprezentacji dostępny jest sposób określania podobieństwa pomiędzy dokumentami i może być on wyrażony za pomocą wartości numerycznej. Wśród metod taksonomicznych szczególną grupę stanowią metody hierarchiczne pozwalające odkryć istniejącą hierarchię w zbiorze dokumentów. Uzyskane w ten sposób wyniki mogą być podstawą do graficznej prezentacji struktury analizowanej kolekcji. Ciekawą pracą dotyczącą porównania metod klasyfikacji wzorcowej dokumentów jest [23].

Drzewa klasyfikacyjne

Celem stosowania drzew klasyfikacyjnych jest identyfikacja reguł klasyfikacji analizowanych obiektów do predefiniowanych klas. Rezultat działania algorytmu prezentowany jest w postaci drzewa decyzyjnego lub w postaci zbioru reguł. Drzewa klasyfikacyjne są często wykorzystywane do wyszukiwania prawidłowości istniejących w bazach i w hurtowniach danych. W przypadku analiz text miningowych tego typu algorytmy będą szczególnie przydatne w przypadku przeprowadzonej wcześniej strukturyzacji informacji pozyskanych z dokumentów.

Sieci neuronowe

Również sieci neuronowe mogą być z powodzeniem wykorzystywane do analizy informacji pochodzących z dokumentów. Bogactwo modeli neuronowych sprawia, że ta grupa metod może być wykorzystana do rozwiązywania zadań z zakresu klasyfikacji wzorcowej i bezwzorcowej oraz do opisu zależności występujących pomiędzy pozyskanymi informacjami. O zastosowaniu sieci Kohonena w text miningu można znaleźć informacje w [11],

zaś o zastosowaniu zmodyfikowanego modelu Kohonena pozwalającego na identyfikację zależności hierarchicznych można przeczytać w [16].

Metoda wektorów nośnych (wspierających, podtrzymujących)

Metoda wektorów nośnych (*SVM - Support Vector Machines*) należy obecnie do najpopularniejszych metod stosowanych w *data miningu*. Określa ona sposób konstrukcji hiperpłaszczyzny rozdzielającej rozpatrywane obiekty, przy czym przy wyznaczaniu granicy pomiędzy obiektami nie uwzględnia się położenia wszystkich obiektów, lecz tylko tych, które znajdują się w najbliższym jej sąsiedztwie (współrzędne tych punktów definiują tzw. wektory podtrzymujące). W przypadku braku możliwości dokonania podziału za pomocą hiperpłaszczyzny dokonuje się zanurzenia obiektów do przestrzeni o większej liczbie wymiarów w taki sposób, aby nowe punkty cechowały się separowalnością liniową.

Do zagadnienia wykorzystania metody SVM w text miningu przywiązuje się szczególną uwagę ze względu na możliwości jej adaptacji do różnych metod reprezentacji informacji tekstowej. SVM jest z powodzeniem wykorzystywana do klasyfikacji wzorcowej dokumentów, zarówno przy stosowaniu reprezentacji za pomocą informacji o częstościach występowania wyrazów ([22], [1]), jak i reprezentacji za pomocą struktur złożonych ([7], [9]).

Analiza powiązań

W analizie informacji pozyskanych z dokumentów może być również przydatna analiza powiązań. Jest ona dogodnym narzędziem identyfikacji prawidłowości w danych ustrukturyzowanych. Wyniki jej działania mogą służyć jako podstawa do wizualizacji istniejących związków ([25]).

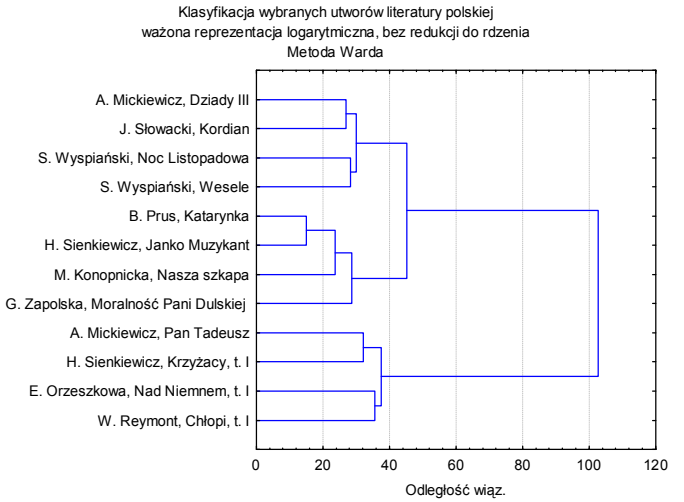
Przykład obliczeniowy

W celu zilustrowania sposobu realizacji analiz text miningowych przedstawiona zostanie procedura klasyfikacji bezwzorcowej tekstów wybranych dzieł literatury polskiej. Analiza obejmowała teksty następujących utworów:

- 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)*.

Do reprezentacji przetwarzanych dokumentów zastosowano metodę bazująca na liście słów. Do realizacji zadania klasyfikacji wykorzystano metody Warda. Klasyfkację uzyskaną w wyniku obliczeń zaprezentowano na dendrogramie przedstawionym na rys. 3.



Rys. 3. Wyniki klasyfikacji wybranych dzieł literatury polskiej

Podsumowanie

Szacuje się, że około 80% informacji przechowywanych jest w postaci dokumentów tekstowych. Fakt ten, w powiązaniu ze stale rosnącą liczbą generowanych dokumentów, stwarza konieczność konstruowania narzędzi wspomagających człowieka w pozyskiwaniu i przetwarzaniu informacji pochodzących z dokumentów tekstowych, w tym również rozwijania metod i narzędzi służących analizie text miningowej.

Z całą pewnością text mining będzie podlegał dalszemu rozwojowi. Dotyczy to metod, narzędzi i zastosowań. Wydaje się, że w niedalekiej przyszłości widoczna będzie coraz mocniejsza integracja text miningu z *data miningiem*, co zaowocuje rozwojem *duo-miningu* ([3]). Należy się również spodziewać wzrostu zainteresowania możliwościami zastosowań text miningu w analizie zawartości serwisów internetowych, a w szczególności systemów biznesu elektronicznego. Podejmowane są również próby zastosowania metod text minigowych w analizie mowy. To wszystko sprawia, że text miningiem warto się zainteresować i w text mining warto zainwestować.

Literatura

1. Basu A., Watters C., Shepherd M., *Support Vector Machines for Text Categorization*, Proceedings of the 36th Hawaii International Conference on System Sciences, 2003.
2. Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D.,. Mercer R.L., Roossin P.S., *A statistical approach to machine translation*,. Computational Linguistics Volume 16, Number 2, June 1990, dostępny w: http://acl.ldc.upenn.edu/J/J90/J90-2002.pdf.
3. Creese G., *Duo-Mining: Combining Data and Text Mining*, DM Review Magazine, 2004, dostępny w: http://www.dmreview.com/article_sub.cfm?articleId=1010449.
4. 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, 41 (6), 391-407, 1990.
5. Ellingsworth M., Sullivan D., *Text Mining Improves Business Intelligence and Predictive Modeling in Insurance*, DM Review Magazine, July 2003, dostępny w: http://www.dmreview.com/article_sub.cfm?articleId=6995.
6. Fan W., Wallace L., Rich S., *Tapping into the Power of Text Mining*, praca przyjęta do publikacji w: The Communications of ACM, dostępny w: http://filebox.vt.edu/users /wfan/paper/text_mining_final_preprint.pdf.
7. Gärtner T., *A survey of kernels for structured data*, ACM SIGKDD Explorations Newsletter archive, Volume 5, Issue 1 (July 2003).
8. Gruber T. R., *A Translation Approach to Portable Ontology Specifications*, dostępny w: http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.
9. Hammer B., Villmann T., *Tutorial: Classication using non-standard metrics*, in: M. Verleysen (ed.), ESANN'2005, to appear.
10. Hearst M.A., *Untangling Text Data Mining*, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper), dostępny w: http://www.sims.berkeley.edu /~hearst/papers/acl99/acl99-tdm.html.
11. Kohonen, T. *Self-organized formation of topologically correct feature maps*. Biological Cy-bernetics, 43, 1982. Springer Verlag, Berlin, Heidelberg, New York.
12. Manning C., Schütze H.,. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
13. Marcotte E.M., Xenarios I., Eisenberg D., *Mining literature for protein-protein interactions*. Bioinformatics 17: 359–363, 2001.
14. Mooney R., Bunescu R., *Mining Knowledge from Text Using Information Extraction*, dostępny w: http://www.cs.utexas.edu/users/ml/papers/text-kddexplore-05.pdf.
15. Nagao, M., *A framework of a mechanical translation between Japanese and English by analogy principle*, In A. Elithorn and R. Banerji (eds.), Artificial and Human Intelligence. Amsterdam: North-Holland, 1984

16. Rauber, A., Dittenbach, M. and Merkl, D., Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural Network Based Approach to Bookshelf Creation and Arrangement, In: Proc. of the 4th European Conference on Research and Advanced Tech-nologies for Digital Libraries (ECDL2000), Springer LNCS 1923, Lisboa, Portugal, 2000.
17. Rebholz-Schuhmann D, Kirsch H, Couto F (2005) *Facts from Text - Is Text Mining Ready to Deliver?* PLoS Biol 3(2); dostępny w: http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0030065.
18. Richardson, S., Dolan W., Menezes A., Pinkham J., *Achieving commercial-quality translation with example-based methods*. In Proceedings of MT Summit VIII, Santiago De Compostela, Spain, pp. 293-298, 2001, dostępny w: http://www.research.microsoft.com/nlp/publications/MTSummit01-Overview-SDR-final.doc.
19. Somers H., *Machine translation: latest developments*. In R. Mitkov (ed.) The Oxford Handbook of Computational Linguistics, Oxford (2003): Oxford University Press, pages 512-528, dostępny w: http://www.co.umist.ac.uk/~harold/Mitkov-book-chapter.pdf.
20. Shull S., *Do You Know What Your Customers are Telling You?*, DM Direct Special Report, May 10, 2005, dostępny w: http://www.dmreview.com/article_sub.cfm?articleID=1027173.
21. Sullivan D., *Integrating Data and Document Warehouses*, DM Review Magazine, July 2001, dostępny w: http://dmreview.com/article_sub.cfm?articleId=3697.
22. Thorsten J., *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, In: Proceedings of ECML-98, 10th European Conference on Machine Learning, edited by Claire Nédellec and Céline Rouveirol, pp 137-142. Springer-Verlag, 1998.
23. Yang Y., Liu X., *A Re-Examination of Text Categorization Methods*, 22nd Annual International SIGIR, dostępny w: http://citeseer.ist.psu.edu/yang99reexamination.html.
24. Yeh A, Hirschman L., Morgan A., *Evaluation of text data mining for database curation*: Lessons learned from the KDD Challenge Cup. Bioinformatics 19: I331-I339, 2003.
25. Wong P.C., Whitney P., Thomas J., *Visualizing Association Rules for Text Mining*, Pacific Northwest National Laboratory, dostępny w: http://www.pnl.gov/infoviz/InfoVis1999Association.pdf.

