POLSKIE STOWARZYSZENIE ZARZĄDZANIA WIEDZĄ
Seria: Studia i Materiały, nr 13, 2008

KAROLINA KULIGOWSKA
MIROSŁAWA LASEK
Katedra Informatyki Gospodarczej i Analiz Ekonomicznych
Uniwersytet Warszawski, Wydział Nauk Ekonomicznych

EKSPLORACJA DANYCH TEKSTOWYCH (TEXT MINING) W PRZEDSIĘBIORSTWIE (TEXT MINING METHODS AND APPLICATIONS IN THE ENTERPRISE)

Streszczenie

Metody eksploracji danych tekstowych Text Mining łączą w sobie techniki Data Mining z analizowaniem treści różnorodnych dokumentów tekstowych. Dzięki tym metodom z nieustrukturyzowanych danych o charakterze tekstowym można odnaleźć nieznane wcześniej informacje oraz dotrzeć do sporej wartości wiedzy dotyczącej przedsiębiorstwa. Pozyskane w ten sposób, usystematyzowane informacje są coraz częściej wykorzystywane do podejmowania decyzji biznesowych.

Text Mining, eksploracja danych tekstowych, Data Mining, Web Mining

1. Wprowadzenie

Od początku badań w dziedzinie sztucznej inteligencji próbowano skonstruować oprogramowanie pozwalające efektywnie analizować dane tekstowe w inteligentny sposób. W miarę jak rozwijano technologię i konstruowano komputery o coraz większej mocy obliczeniowej, możliwe było przeprowadzanie coraz bardziej zaawansowanych analiz danych. Dzięki aplikacjom analitycznym programy zaczęły przetwarzać obszerne bazy danych cyfrowych o wiele efektywniej i szybciej niż człowiek. Pomimo to maszyny wciąż nie radziły sobie z podstawową umiejętnością ludzką: rozumieniem i przetwarzaniem komunikatów w języku naturalnym.

Dalsze badania naukowe prowadzone w zakresie lingwistyki obliczeniowej (ang. computational linguistics) okazały się na tyle owocne, że zaczęto wytwarzać oprogramowanie do analiz tekstu, tzw. Text Mining. Narzędzia Text Miningu stanowią połączenie metod Data Miningu zastosowanych do przetwarzania języka naturalnego. Narzędzia te umożliwiają wyłuskiwanie cennych informacji z bardzo wielu, różnorodnych dokumentów tekstowych, a co za tym idzie – odkrywanie nieznanych wcześniej współzależności między danymi oraz powiązań pomiędzy dokumentami (Gaizauskas, 2004). Badania nad Text Miningowymi metodami eksploracji danych wydają się być bardzo obiecujące, gdyż pozwalają na zaoszczędzenie czasu i pieniędzy, które musiałyby zostać przeznaczone na przeczytanie i ewentualne eksplorowanie przez człowieka ogromnego repozytorium dokumentów tekstowych. Text Mining jest już stosowany w przedsiębiorstwach, a niektóre z zastosowań zostaną omówione w niniejszym artykule.

2. Metody eksploracji danych tekstowych

Podczas używania narzędzi bazujących na Data Miningu informacje zostają wyłuskiwane z ustrukturyzowanych baz danych. W procesie Text Miningu natomiast dane są wydobywane z nieustrukturyzowanych treści dokumentów tekstowych zapisanych w języku naturalnym.

W celu przeprowadzenia analizy tekstu dokument powinien być na samym początku przekształcony w odpowiednią formę. Etap ten nazywa się wstępną obróbką pliku tekstowego (ang. preprocessing), podczas którego dane tekstowe zapisane w różnorodnych formatach zostają zaimportowane do pojedynczego zbioru, łatwego do późniejszego odczytywania.

Kiedy „surowe" dokumenty są już przekształcone w jednolity format kodowania, tekst jest przeszukiwany i następnie poddawany rozbiorowi. Jest to kolejny krok w procesie analizy kolekcji dokumentów. Podczas rozbioru struktury dokumentu (ang. parsing) z dalszej analizy wyłączone zostają wyrazy o niskiej wartości informacyjnej. Rodzajniki, spójniki, przyimki i inne nieistotne semantycznie części mowy zgromadzone są na osobnej liście, tzw. stop liście (ang. stop list), za pomocą której można kontrolować pomijane wyrazy. W etapie tym zostają również wyodrębnione terminy, które mają istotne znaczenie i które należy włączyć do analizy; mogą to być pojedyncze wyrazy, wyrażenia, nazwy oraz numery.

Podczas przeszukiwania dokumentu następuje także automatyczne znajdowanie podstawy fleksyjnej (ang. stemming), czyli sprowadzenie wszystkich odmian i form danego wyrazu do jego formy podstawowej. Czynność ta ułatwia analizę terminów bardziej abstrakcyjnych, choć oczywiście istnieją terminy mające identyczne znaczenie kontekstowe, lecz oparte na innych podstawach fleksyjnych. W takim przypadku niezbędna jest lista synonimów, które mają takie samo znaczenie, choć nie wywodzą się bezpośrednio z tych samych form podstawowych (Węcel, 2005).

Efektem przeszukiwania dokumentu jest wygenerowanie liczbowej reprezentacji danego dokumentu. Może ona być oparta na prostych metodach statystycznych bazujących na częstości i współwystępowaniu wyrazów. W celu analizy liczby wyrazów w zbiorze dokumentów najczęściej tworzona jest macierz częstości występowania wyrazów w dokumencie. Wagi używane do mierzenia częstości występowania słów zależą od częstości występowania danego wyrazu w dokumencie oraz w kolekcji dokumentów jako całości. Po zmierzeniu częstości można następnie stosować filtrowanie tekstu i ekstrakcja faktów.

Celem stosowania eksploracji danych tekstowych jest przeszukanie dokumentów zawierających nieustrukturyzowany tekst, wydobycie z niego wartościowych dla analizy słów, do których następnie stosuje się różne algorytmy Data Miningu. Wydobyte informacje mogą być użyte do sporządzania streszczeń dokumentów, określenia podobieństw pomiędzy wieloma dokumentami, znalezienia zależności pomiędzy jednostkami tekstu lub tworzenia rankingów dokumentów (Filipowska, 2004). Widać więc, że metody Text Miningu są potężnym narzędziem przekształcającym bezładny tekst w liczby, które są łatwiejsze do analitycznej obróbki i mogą być następnie włączone do analiz, takich jak modelowanie Data Miningowe, predykcja lub biznesowe zastosowania klasteryzacji i klasyfikacji.

3. Text Mining w przedsiębiorstwie

Głównym zadaniem metod eksploracji tekstu jest wyłuskiwanie istotnych danych i użycie ich do sporządzania prognoz i podejmowania decyzji biznesowych. Do osiągnięcia tego celu analitycy wykorzystują klasteryzację dokumentów oraz ich klasyfikowanie.

Klasteryzacja kolekcji dokumentów umożliwia sporządzenie ich streszczeń bez konieczności czytania przez człowieka każdego dokumentu z osobna. Klaster zawierający kilka tysięcy dokumentów może pomóc w ujawnieniu ważnych zagadnień i kluczowych idei związanych z funkcjonowaniem przedsiębiorstwa, a zawartych w zgromadzonych w firmie dokumentach. Klasteryzację dokumentów stosuje się w analizie danych ankietowych, analizie opinii klientów lub zbiorów wiadomości e-mail do odkrycia wcześniej nieznanej wiedzy. Klasteryzacja daje również wskazówki jakie wyrazy mają tendencję do bycia używanymi łącznie lub jakie kategorie słownictwa występują w kolekcji dokumentów.

Klasyfikowanie dokumentów polega na ich rozdzieleniu pomiędzy wcześniej zdefiniowane kategorie. Można powiedzieć, że klasyfikowanie jest w zasadzie formą predykcji. Jest ona często używana do inteligentnego filtrowania wiadomości e-mail lub automatycznego wykrywania spamu.

Najbardziej spektakularne i obiecujące zastosowania Text Miningu dotyczą sporządzania prognoz w takich dziedzinach, jak: giełda i kursy walut, ocena satysfakcji klienta oraz przewidywanie zachowań i preferencji klienta (Weiss, 2005). Inne typowe obszary zastosowań eksploracji danych tekstowych:

- zmiany cen akcji na giełdzie przewidziane na podstawie prasowych informacji o kondycjach finansowych firm;
- koszty usług prognozowane na podstawie opisu problemu;
- identyfikacja konkretnych słów i wyrażeń dla procesu filtrowania wiadomości e-mail w celu wykrycia spamu;
- satysfakcja konsumenta przewidziana na podstawie analiz danych ankietowych oraz komentarzy klientów wpisanych na stronie internetowej;
- zbadanie próbek artykułów napisanych przez jedną osobę może być podstawą do udowodnienia jej autorstwa innego fragmentu tekstu, który ma kilku potencjalnych autorów.

Inne zastosowania Text Miningu mogą dotyczyć analiz ankiet złożonych z pytań otwartych, automatycznego przetwarzania wiadomości, analiz roszczeń ubezpieczeniowych oraz analiz różnorodnych diagnoz (Hearst, 1999).

4. Eksploracja danych zawartych w internecie (Web Mining) na potrzeby przedsiębiorstw

Internet w bardzo intensywny sposób oddziałuje na współczesne społeczeństwo, zmieniając sposoby wymiany informacji oraz zbierania danych. To Internet jest uważany za najobszerniejsze źródło informacji na całej planecie. Można go określić jako niewiarygodnie wielki magazyn wszelakich nieuporządkowanych danych. Nic więc dziwnego, że również w internecie zaczęto stosować techniki Text Minigu do danych zawartych w internecie, czyli po prostu Web Miningu.

Narzędzia eksploracji danych internetowych umożliwiają przeszukiwanie danych rozproszonych w całej światowej sieci internetowej. Rozróżnia się trzy rodzaje Web Miningowych analiz danych, mianowicie: eksploracja zawartości stron internetowych (ang. Web content mining), eksploracja struktur internetowych (ang. Web structure mining) oraz eksploracja użytkowania internetu (ang. Web usage mining). Pierwsza metoda skupia się na wyszukiwaniu użytecznych informacji bezpośrednio z zawartości stron internetowych i dokumentów

zamieszczonych w internecie. Druga metoda umożliwia odkrywanie modeli struktur hiperłączy. Trzecie podejście odnosi się do technik przewidywania zachowania użytkowników na podstawie ich wcześniej zaobserwowanych wzorców zachowań (Wang, 2000).

Internauci powszechnie już używają w codziennej pracy narzędzi, takich jak wyszukiwarki, gdyż zależy im na szybkim i precyzyjnym odnalezieniu ważnych informacji. Z drugiej strony dostawcy internetu starają się przewidzieć zachowania użytkowników oraz wzorce ich nawigacji w sieci w celu zredukowania przeciążenia w ładowaniu stron oraz w celu personalizacji dostarczanych informacji. Analitycy w firmie szczególnie cenią sobie zrozumienie i możliwości predykcji preferencji i oczekiwań użytkownika. Wszystkie wyżej wymienione grupy chciałyby używać odpowiednich narzędzi Web Miningowych, które pomogłyby im rozwiązać problemy dotyczące ogromnej ilości danych zawartych w internecie.

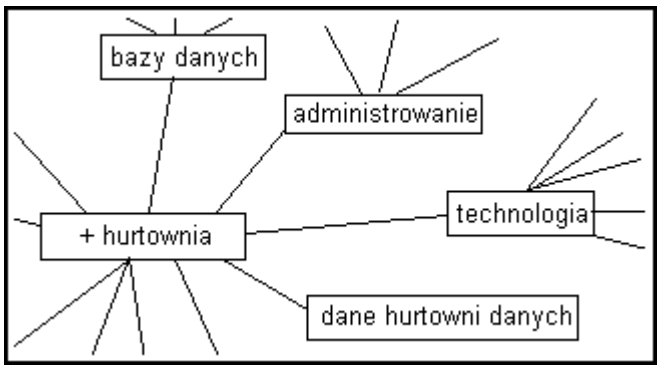
Metody Web Miningu przyczyniają się do sprawniejszego wykonywania zadań wewnątrz przedsiębiorstwa poprzez efektywną eksplorację portali internetowych. Narzędzia te umożliwiają dokładne personalizowanie serwisów internetowych poprzez śledzenie wzorców nawigacji użytkowników oraz na tej podstawie odpowiednią indywidualizację zawartości stron. Wykorzystanie wiedzy dotyczącej klientów oraz ich preferencji jest bardzo ważnym aspektem wykorzystywanym przy podejmowania decyzji rynkowych w przedsiębiorstwie. Dzięki Web Miningowi satysfakcja klienta może być mierzona i analizowana już choćby na bazie wypełnianych on-line kwestionariuszy (Night, 1999).

Przyspieszony wzrost źródeł informacji dostępnych w internecie oraz zainteresowanie handlem elektronicznym sprawia, że globalna sieć stała się bardzo atrakcyjnym miejscem wymiany doświadczeń naukowców i praktyków biznesu dzięki możliwości sprawnego przeszukiwania informacji oraz metodom sztucznej inteligencji, a szczególnie przetwarzania języka naturalnego.

5. Narzędzia wykorzystywane do Text Miningu

Pomiędzy wieloma dostępnymi programami używanymi do Text Miningu najczęściej używanymi są: Text Miner (SAS), Text Miner (StatSoft), Text Analyst (Megaputer Intelligence) oraz Text Mining Solutions (NetOwl). Narzędzia te potrafią zanalizować tekst znajdujący się w kolekcji dokumentów, a następnie dokonać na tej podstawie predykcji lub dalszej eksploracji tekstu. Powyższe oprogramowanie ułatwia także nawigację w bardzo złożonych bazach danych oraz umożliwia sporządzenie streszczeń bez konieczności zapoznawania się z całymi dokumentami. Dodatkowe opcje, w jakie wyposażone są te programy to klasteryzacja dokumentów, automatyczne rozpoznawanie złożonych wyrażeń oraz znajdywanie grup słów o podobnym znaczeniu lub znajdywanie grup podobnych treściowo dokumentów.

Używając modułu SAS Text Miner można także analizować powiązania pomiędzy terminami. Przykładową graficzną prezentację słów występujących najczęściej razem w tekście można obejrzeć na rysunku poniżej.



Rys. 1. Słowa najczęściej występujące z wyrazem „hurtownia"
(źródło: opracowanie własne na podstawie programu SAS Text Miner)

6. Literatura

Filipowska A., *Jak zaoszczędzić na czytaniu? Automatyczne tworzenie abstraktów z dokumentów*, *Gazeta IT*, nr 3 (22), 2004, 1-6.

Gaizauskas R., Saggion H., *Multi-Document Summarization by Cluster/Profile Relevance and Redundancy Removal*, *Proceedings of the HLT/NAACL Document Understanding Workshop*, Boston 2004, 1-8.

Hearst M., A., *Untangling Text Data Mining*, *Proceedings of ACL, 37th Annual Meeting of the ACL*, New Jersey 1999, 3-10.

Night K., Mining Online Text, *Communications of the ACM 42(11), ACM Press*, New York 1999, 58–61.

Wang Y., *Web Mining and Knowledge Discovery of Usage Patterns*, *CS 748T Project 2000*, 1-25

Weiss S. (red.), *Text Mining: Predictive Methods for Analyzing Unstructured Information*, *Springer Science and Business Media*, New York 2005.

Węcel K., *SAS, rejestry państwowe i text mining*, *Gazeta IT*, nr 9 (39), 2005, 1-4

Abstract

Text Mining methods consist of Data Mining algorithms applied to unstructured textual data. Those methods allow to explore quickly thousands of documents and to extract previously unktnown patterns and correlations used in business decision making and other managerial activities in the enterprise.

Keywords: Text Mining, eksploracja danych tekstowych, Data Mining, Web Mining