ROZDZIAŁ XX

METODY SZTUCZNEJ INTELIGENCJI W PRZEWIDYWANIU WARTOŚCI INDEKSU GIEŁDOWEGO Z WYKORZYSTANIEM ARTYKUŁÓW PRASOWYCH

Mateusz KOBOS*, Jacek MAŃDZIUK*

1. Wprowadzenie

Zarządzanie wiedzą i informacją w przedsiębiorstwie wymaga łączenia i analizowania danych pochodzących z różnych źródeł. Dąży się do tego, by przeprowadzana analiza danych była wykonywana w miarę możliwości automatycznie. Jednocześnie, wymaga się, by tego typu analiza dawała na tyle wiarygodne wyniki, by można było je wykorzystać do rozwiązywania praktycznych problemów.

Przykładem systemów, które starają się realizować powyższe wymagania, są automatyczne systemy przewidujące wartości indeksu giełdowego. Część z tych systemów korzysta z dwóch podstawowych źródeł informacji: numerycznego szeregu czasowego notowań akcji lub indeksu giełdowego oraz danych tekstowych, na które składają się artykuły opublikowane w specjalistycznej prasie. Tego typu systemy predykcyjne, wywodzące się z dziedziny sztucznej inteligencji, zostały omówione w niniejszym rozdziale. W podrozdziale 2 omówiono różne rodzaje systemów predykcyjnych i ich działanie. Następnie, w podrozdziale 3 zaprezentowano wyniki wstępnych badań. Natomiast w podrozdziale 4 zostało przedstawione podsumowanie całego rozdziału.

2. Działanie systemu predykcyjnego

Ogólny schemat systemu predykcyjnego korzystającego zarówno z notowań akcji lub indeksu giełdowego jak i artykułów prasowych jest przedstawiony poniżej. Sercem systemu jest algorytm predykcyjny (aproksymator lub klasyfikator) a danymi wejściowymi algorytmu są artykuły prasowe i szereg czasowy notowań. Przed użyciem w systemie, dane te są poddawane wstępnej obróbce (ang. *preprocessing*) polegającej np. na zamianie szeregu notowań na szereg *zwrotów* - względnych różnic, usunięciu błędnych czy uzupełnieniu brakujących danych. Dane tekstowe są zamieniane na reprezentację numeryczną, która może

* Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych, e-mail: M.Kobos@mini.pw.edu.pl, J.Mandziuk@mini.pw.edu.pl

być użyta bezpośrednio przez algorytm predykcyjny. Wynikiem działania algorytmu predykcyjnego jest przewidywana wartość akcji/indeksu w przypadku aproksymatora lub kategoria przewidywanej wartości (wzrost, spadek, brak zmiany notowanych aktywów) w przypadku klasyfikatora. Taka prognoza może być w dalszej kolejności wykorzystana przez algorytm inwestycyjny wspomagający inwestora lub dokonujący samodzielnie inwestycji na rynku.

2.1. Rodzaje algorytmów predykcyjnych

W literaturze występuje wiele rodzajów algorytmów predykcyjnych. Można je zgrubnie podzielić na dwie następujące klasy:

1. algorytmy wykorzystujące metody sztucznej inteligencji,
2. algorytmy oparte na modelach matematycznych.

Do najczęściej stosowanych metod sztucznej inteligencji należą: sieci neuronowe (np. perceptron wielowarstwowy, sieć probabilistyczna) [6, 4], Support Vector Machines [7, 8], algorytmy genetyczne [11], drzewa decyzyjne, reguły decyzyjne [9], naiwne modele Bayes'a [3], k-Nearest Neighbours [8]. Wśród metod należących do drugiej grupy szczególnie popularne są modele przewidywania wartości szeregów czasowych wywodzące się z modeli ARIMA i GARCH [2] oraz Linear Discriminant Analysis [1].

2.2. Dane wejściowe

Jak wspomnieliśmy w części wprowadzającej, dane wejściowe wykorzystywane przez algorytm predykcyjny są dwojakiego rodzaju:

1. dane tekstowe,
2. szeregi czasowe.

Najczęściej stosowane dane tekstowe pochodzą z popularnych czasopism finansowych („The Wall Street Journal", „Financial Times"), komunikatów prasowych firm oraz komunikatów prasowych pochodzących z agencji informacyjnych np. „Dow Jones Newswire", „Reuters" czy „Bloomberg". Do najpopularniejszych szeregów czasowych należą notowania akcji spółek, notowania indeksów giełdowych, kursy wymiany walut, zwroty z obligacji korporacyjnych.

2.3. Numeryczne reprezentacje tekstu

Numeryczne reprezentacje tekstu, za pomocą których są przedstawiane artykuły prasowe, można podzielić na dwie grupy:

1. reprezentacje generowane automatycznie,
2. reprezentacje korzystające z wiedzy *a priori.*

Wśród reprezentacji generowanych automatycznie najpopularniejsza jest reprezentacja „bag-of-words" zwana też „vector space". W reprezentacji tej każdemu dokumentowi odpowiada wektor numeryczny. Każda ze współrzędnych wektora odpowiada jednemu słowu występującemu w zbiorze wszystkich dokumentów. Wartość, jaką przyjmuje współrzędna wektora, wskazuje jak ważne jest dane słowo w dokumencie odpowiadającym wektorowi. Mimo swej prostoty, reprezentacja „bag-of-words" dobrze sprawdza się w praktyce. Reprezentacja ta zostanie dokładniej przedstawiona w dalszej części rozdziału.

Wśród reprezentacji korzystających z wiedzy *a priori* można rozróżnić następujące podejścia: analiza artykułu przez człowieka i zastosowanie predefiniowanych reguł eksperckich dotyczących treści artykułu [6], tworzenie wektora słów kluczowych używanych w reprezentacji „bag-of-words" z predefiniowanych słów kluczowych [9, 8], wykorzystanie

wiedzy eksperckiej zapisanej za pomocą tzw. „cognitive map” [4], rozpoznawanie za pomocą wyrażeń regularnych przynależności do jednej z predefiniowanych kategorii [11].

2.3.1. Reprezentacja tekstu typu „bag-of-words”

Poniżej przedstawimy reprezentację tekstu typu „bag-of-words”, którą zastosowaliśmy w przeprowadzonych eksperymentach wstępnych. Przez $a(D,w)$ będziemy oznaczali wartość, jaką przyjmuje współrzędna wektora odpowiadająca słowu w w dokumencie D, przez T będziemy oznaczali zbiór uczący dokumentów, który służy do nauki algorytmu predykcyjnego, czyli do dobrania parametrów algorytmu.

Do najpopularniejszych reprezentacji typu „bag-of-words” można zaliczyć:

- reprezentację binarną: $a(D,w)=\begin{cases}1 & w\in D\\ 0 & w\notin D\end{cases}$,
- reprezentację TF: $a(D,w)=TF(D,w)$,
- reprezentację TF-IDF: $a(D,w)=TF(D,w)IDF(T,w)$,

w których

$TF(D,w)$ - częstość występowania słowa w w dokumencie D,

$IDF(T,w)=\log(\frac{|T|}{|T_w|})$ - logarytm z odwrotności częstości występowania dokumentów zawierających słowo w wśród dokumentów ze zbioru T,

T_w - zbiór dokumentów zawierających słowo w (podzbiór T).

Słowo, któremu przypisana jest duża wartość współrzędnej wektora odpowiadającego dokumentowi, można interpretować jako słowo istotne. Natomiast słowo, któremu jest przypisana wartość bliska zeru, można interpretować jako nieistotne. Przy tej interpretacji, reprezentacja TF wskazuje jako ważne te słowa, które często występują w dokumencie. Z kolei reprezentacja TF-IDF jako ważne wskazuje te słowa, które często występują w danym dokumencie, ale jednocześnie rzadko występują w innych dokumentach.

2.3.2. Słownik w reprezentacji tekstu typu „bag-of-words”

Zbiór słów, którym odpowiadają współrzędne wektorów w reprezentacji „bag-of-words”, nazywamy *słownikiem*. Początkowo słownik składa się ze wszystkich słów występujących we wszystkich dokumentach. W trakcie wstępnej obróbki danych część z tych słów jest usuwana. Do usuwania słów uznanych za zbędne stosuje się najczęściej poniższe metody redukcji rozmiaru słownika.

- Metody związane z cechami lingwistycznymi słów:
  - odrzucenie „stop words” (przyimki, zaimki, spójniki),
  - utożsamienie synonimów,
  - *stemming* – utożsamienie pokrewnych słowotwórczo słów i sprowadzenie ich do jednego rdzenia.
- Metody związane z cechami całego zbioru dokumentów:
  - odrzucenie najrzadziej lub najczęściej występujących słów,
  - odrzucenie słów występujących w podobnej liczbie w każdej z predefiniowanych klas dokumentów (słowa o dużej entropii informacji),
  - odrzucenie słów o najmniejszym współczynniku CTF-IDF, który dla danego słowa w jest określony za pomocą wzoru: $CTF-IDF(w)=\frac{\text{count}(w)}{|T_w|}$, gdzie

count(w) jest liczbą wystąpień słowa w we wszystkich dokumentach. Można zauważyć, że jako słowa istotne (którym odpowiada wysoki współczynnik CTF-IDF) są oznaczane te, które występują często, ale w małej liczbie dokumentów.

3. Wyniki badań

W niniejszej części zostały opisane wyniki wstępnych eksperymentów przeprowadzonych przez autorów. Badania miały na celu sprawdzenie możliwości przewidywania wartości indeksu giełdowego na podstawie szeregu czasowego notowań wybranego indeksu giełdowego oraz artykułów prasowych.

3.1. Dane wejściowe

W badaniach zostały wykorzystane streszczenia artykułów prasowych z przedziału czasowego: 2006.04.01-2007.04.01 oraz szereg czasowy zwrotów z notowań indeksu z tego samego przedziału czasowego.

W badaniach wykorzystano łącznie 46623 artykuły z czasopisma „The Wall Street Journal". Artykuły zostały udostępnione przez wydawcę „The Wall Street Journal" i dystrybutora czasopisma – firmę „ProQuest". Podstawowe wskaźniki statystyczne opisujące rozkład liczby dziennie publikowanych artykułów są następujące: minimum = 1; maksimum = 180, średnia = 104; standardowe odchylenie = 48. W badanym przedziale czasowym występuje 5 dni, z którymi związana jest mała (mniejsza niż 6) liczba opublikowanych artykułów (anomalia ta jest obecna w oryginalnym źródle danych).

Szereg czasowy zwrotów opisuje zwroty z notowań indeksu S&P500. Dane pochodzą z serwisu http://finance.yahoo.com. Podstawowe wskaźniki statystyczne opisujące rozkład zwrotów są następujące: minimum = -0,038; maksimum = 0,022; średnia = 0; odchylenie standardowe = 0,007.

3.2. Algorytm predykcyjny i reprezentacja tekstu

Jako algorytm predykcyjny została zastosowana sieć neuronowa typu perceptron wielowarstwowy z zaimplementowanym algorytmem uczenia RProp. Sieć posiadała jedną warstwę ukrytą zawierającą liczbę neuronów równą połowie liczby neuronów znajdujących się w warstwie wejściowej, natomiast w warstwie wyjściowej znajdował się jeden neuron. W trakcie uczenia wykorzystano technikę „early stopping", w której zbiór walidacyjny stanowiło 20% losowo wybranych wektorów ze zbioru uczącego.

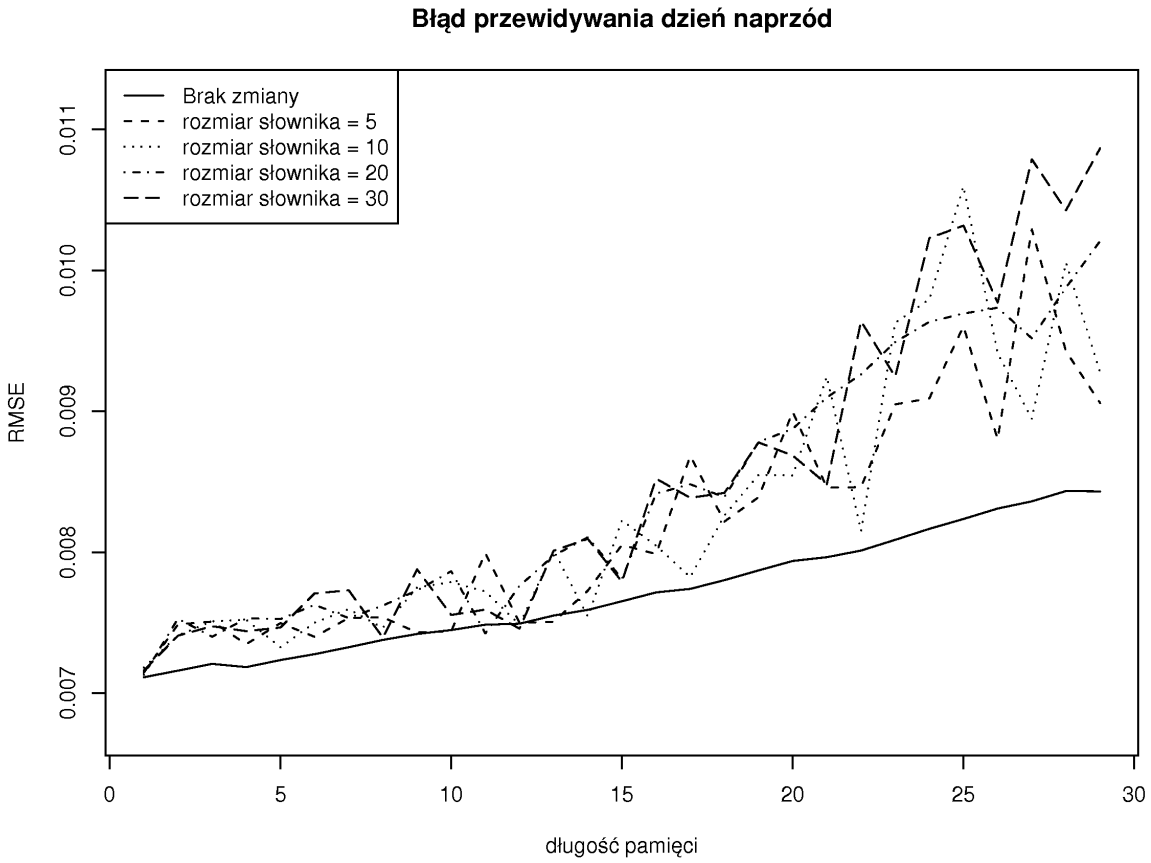
Do reprezentacji tekstu wybrano metodę TF-IDF. Ze słownika wyrzucono „stop words" i zastosowano algorytm Porter'a [10] do wykonania stemming-u. W celu dalszej redukcji rozmiaru słownika, zastosowano metodę polegającą na odrzuceniu słów o małym współczynniku CTF-IDF.

Szereg czasowy podzielono w 70% jego długości na dwie części. Pierwsza z nich służyła do nauki algorytmu predykcyjnego, a druga wyłącznie do testowania algorytmu i obliczania błędu predykcji.

3.3. Opis badań

W trakcie eksperymentów sprawdzono m.in. zależność między „długością pamięci" algorytmu predykcyjnego a błędem przewidywania. Przez *długość pamięci* rozumiemy tu liczbę poprzednich dni, z których udostępniane są algorytmowi predykcyjnemu wcześniej

wspomniane dane (tekstowe i numeryczne) mające pomóc w przewidzeniu zwrotu ceny akcji z kolejnego dnia. Eksperymenty przeprowadzono dla słowników o różnych rozmiarach. Wyniki porównano z prostą regułą przewidującą, która mówi, że jutrzejsza wartość indeksu będzie taka sama jak dzisiejsza, a więc zwrot będzie zawsze równy zero. Należy zauważyć, że jest to dość silna heurystyka wynikająca z modelu błądzenia losowego (ang. *random walk model*) związanego z hipotezą rynku efektywnego. Reguła ta jest często stosowana do sprawdzania efektywności przewidywania finansowych szeregów czasowych. Jako miarę błędu przewidywania wykorzystano pierwiastek z błędu średniokwadratowego (*RMSE*).



Rys. 5.1. Zależność między długością pamięci a błędem przewidywania zwrotu z notowania indeksu dzień naprzód. Ciągła linia odpowiada regule mówiącej, że jutrzejsza wartość indeksu będzie taka sama jak dzisiejsza.

3.4. Analiza otrzymanych wyników i możliwe rozszerzenia systemu

Na podstawie przeprowadzonych badań otrzymano następujące wnioski. Algorytm predykcyjny osiąga gorsze wyniki, niż heurystyka zakładająca utrzymanie wartości w kolejnym dniu, a wraz z długością pamięci błąd przewidywania rośnie.

Główną przyczyną uzyskania niezadowalających wyników jest najprawdopodobniej problem z dobraniem odpowiednich słów do słownika. Zastosowany współczynnik CTF-IDF w przypadku użytych danych sprawia, że do słownika wybierane są głównie słowa mało znaczące ale popularne, które występują we wszystkich dokumentach. Powoduje to, że po zastosowaniu reprezentacji TF-IDF, większość wektorów odpowiadających dokumentom jest zerowa, a w rezultacie, zgodnie z tą reprezentacją, nie ma znaczenia. Przykładem tego typu mało znaczących słów mogą być słowa, które znalazły się w słowniku o rozmiarze 10: „in”, „on”, „Mr”, „sai”, „he”, „lear”, „compani”, „U.S”, „hi”, „Nuv”. Aby rozwiązać ten problem,

planowane są prace nad doborem bardziej adekwatnego współczynnika mierzącego ważność słowa wśród wszystkich dokumentów. Do poprawy wyników może przyczynić się również lepsza wstępna obróbka słów, dzięki której do słownika nie trafią słowa typu: „in", „on", „he".

Przyczyną niskiej jakości przewidywania może być też zbyt prosta reprezentacja tekstu. W związku z tym, planuje się badać występowanie całych predefiniowanych zwrotów zamiast pojedynczych słów. Występowanie zwrotów mierzone byłoby w sposób przybliżony, tzn. uznawalibyśmy, że dany zwrot występuje w dokumencie jeśli np. dwa słowa tworzące zwrot występowałyby odpowiednio blisko siebie w tekście.

Kolejną przyczyną niepowodzenia jest najprawdopodobniej nieuwzględnienie wydźwięku danego artykułu. Wydaje się, że uwzględnienie informacji o tym czy artykuł ma zabarwienie pozytywne, negatywne czy też jest neutralny mogłaby znacznie zwiększyć skuteczność rozważanego systemu. W związku z tym, planowana jest automatyczna klasyfikacja dokumentu według jego wydźwięku. Aby osiągnąć ten cel, zostanie podjęta próba wyszukiwania w dokumencie wcześniej zdefiniowanych zwrotów świadczących o pozytywnej lub negatywnej opinii zawartej w artykule. Przykładowymi zwrotami tego typu mogą być: „investors are anxious", „markets values decrease", „financial problems", „good financial results", „bull market", itp.

Na polepszenie wyników przewidywania może mieć również wpływ uwzględnienie innych niż streszczenie części artykułów. Planuje się sprawdzenie wydajności przewidywania przy użyciu tytułu i pierwszego akapitu tekstu artykułu. Te części artykułu często zawierają podsumowanie całej publikacji, a więc są szczególnie istotne dla zadania rozumienia zawartości tekstu. W celu polepszenia wyników zostaną również przypisane większe wagi (odpowiadające większej istotności) słowom występującym w tytule niż słowom znajdującym się w streszczeniu i pierwszym akapicie tekstu artykułu.

Ponadto planuje się przewidywanie cen akcji danej firmy lub firm pochodzących z określonego segmentu rynku na podstawie artykułów związanych z daną firmą lub firmami z danego segmentu rynku. Wyniki przewidywania dla tego wariantu mogą być potencjalnie lepsze od wyników przewidywania wartości indeksu ze względu na prawdopodobnie większy związek między artykułem na temat danej firmy a ceną akcji firmy.

W przypadku przewidywania cen akcji firmy, pozytywny wpływ na jakość predykcji może mieć też uwzględnienie tematyki danego artykułu – np. artykuł opisujący sprawozdanie finansowe firmy może mieć większy wpływ na ceny akcji od opisu nowowprowadzonego na rynek produktu tej firmy. Można tu wykorzystać predefiniowaną klasyfikację tematyczną wprowadzoną przez dystrybutora czasopisma lub wyszukiwać w artykule słów kluczowych związanych z daną tematyką. Zbliżone rozwiązanie przy analizie tematycznej artykułów naukowych było wykorzystane w pracy [5].

Ciekawe i warte sprawdzenia wydaje się również porównanie wyników otrzymanych przy użyciu słownika generowanego automatycznie z wynikami otrzymanymi przy użyciu słownika przygotowanego przez człowieka, który zawierałby słowa lub zwroty uznane jako ważne dla oceny znaczenia dokumentu. W przypadku przewidywania zwrotów z akcji firm danej branży można by zastosować słownik słów związanych z daną branżą, np. dla branży komputerowej istotnymi zwrotami mogą być: „security flaw", „bug", „hacker attack" i inne.

4. Podsumowanie

W niniejszym rozdziale przedstawiono sposób budowy i działania systemów służących do automatycznego przewidywania notowań akcji czy indeksów giełdowych. Przedstawiono popularne rodzaje reprezentacji tekstu typu „bag-of-words". Zaprezentowano wyniki wstępnych eksperymentów z tej dziedziny oraz zaproponowano dalsze ścieżki rozwoju przedstawionego systemu.

5. Podziękowania

Autorzy dziękują firmie ProQuest i wydawcy „The Wall Street Journal" za udostępnienie danych tekstowych wykorzystanych podczas wstępnych eksperymentów.

Bibliografia

[1] Aasheim C., Kohler G.J.: Scanning World Wide Web documents with the vector space model, *Decision Support Systems*, vol. 42, issue 2, pp. 690-699, 2006.
[2] Edmonds R.G., Kutan A.M.: Is public information really irrelevant in explaining asset returns?, *Economics Letters*, vol. 76, issue 2 , pp. 223-229, 2002.
[3] Gidófalvi G., Elkan. C.: *Using News Articles to Predict Stock Price Movements*, (Technical Report). Department of Computer Science and Engineering, University of California, San Diego (USA), 2003.
[4] Hong T., Han I.: Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks, *Expert systems with applications*, vol. 23, pp. 1-8, 2002.
[5] Klincewicz K., Miyazaki K.: Software Sectoral Systems of Innovation in Asia. Empirical Analysis of Industry-Academia Relations. *Proceeding of the IEEE International Conference on Management of Innovation and Technology*, Singapore, 2006, vol. 1, pp. 494-498, IEEE, Singapore (China), 2006
[6] Kohara K., Ishikawa T., Fukuhara Y., Nakamura Y.: Stock price prediction using prior knowledge and Neural Networks, *Intelligent systems in accounting, finance and management*, vol. 6, pp. 11-22, 1997.
[7] Mittermayer M.-A.: Forecasting Intraday Stock Price Trends with Text Mining Techniques, *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04)*, Hawaii, vol. 3, IEEE Computer Society, Washington, DC (USA), 2004.
[8] Mittermayer M.-A., Knolmayer G.F.: NewsCATS: A News Categorization And Trading System, *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong (China), pp. 1002-1007, IEEE Computer Society , Washington, DC (USA), 2006.
[9] Peramunetilleke D., Wong R.K.: Currency Exchange Rate Forecasting from News Headlines, *Proceedings of the 13th Australasian database conference (ADC2002)*, Melbourne, vol. 5, pp. 131-139, Australian Computer Society, Melbourne (Australia), 2002.
[10] Porter M.F.: An algorithm for suffix stripping, *Program*, vol. 14(3), pp. 130−137, 1980.
[11] Thomas D.: News and Trading Rules (Ph.D. thesis). Carnegie Mellon University, 2003.
[12] Wuthrich B., Cho V., Leung S., Permunetilleke D., Sankaran K., Zhang J., Lam W., Daily Stock Market Forecast from Textual Web Data, *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, San Diego, CA, (USA), vol. 3, pp. 2720-2725, 1998.

{Odrębna strona streszczeń i adresów}

Metody sztucznej inteligencji w przewidywaniu wartości indeksu giełdowego z wykorzystaniem artykułów prasowych

M. Kobos*, J. Mańdziuk**

*Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych,
M.Kobos@mini.pw.edu.pl

** Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych,
J.Mandziuk@mini.pw.edu.pl

Rozdział 5. Metody sztucznej inteligencji w przewidywaniu wartości indeksu giełdowego z wykorzystaniem artykułów prasowych. Analiza danych jest ważną częścią zarządzania informacją i wiedzą w przedsiębiorstwie. W niniejszym rozdziale przedstawiamy przegląd literaturowy automatycznych systemów wywodzących się z dziedziny sztucznej inteligencji służących do przewidywania wartości indeksu giełdowego przy wykorzystaniu numerycznych szeregów czasowych oraz artykułów prasowych. Zaprezentowane zostało również wstępne podejście do problemu implementacji tego typu systemu.

Artificial Intelligence methods in prediction of stock index values with usage of newspaper articles

M. Kobos*, J. Mańdziuk**

*Warsaw University of Technology, Faculty of Mathematics and Information Science,
M.Kobos@mini.pw.edu.pl

**Warsaw University of Technology, Faculty of Mathematics and Information Science,
J.Mandziuk@mini.pw.edu.pl

Chapter 5. Artificial Intelligence methods in prediction of stock index values with usage of newspaper articles. Analysis of various kinds of data is an important part of information and knowledge management in a company. This chapter contains an overview of literature concerning Artificial Intelligence automatic prediction systems applied to prediction of stock index values with usage of numerical time series and newspaper articles. Moreover, an initial approach to implementation of such a system is proposed.