PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1081 — 2005
Nowoczesne technologie informacyjne w zarządzaniu

Klaudia Smoląg, Waldemar Jędrzejczyk

EKSPLORACJA DANYCH Z WYKORZYSTANIEM PAKIETU SPHINX

1. Wprowadzenie

Wraz z rozwojem technologii informatycznej obserwujemy wzrost ilości gromadzonych danych. Wielość danych i informacji zgromadzona w różnych systemach informatycznych utrudnia analizę danych oraz podejmowanie na ich podstawie decyzji. Z kolei wśród ogromnej liczby pozornie nieprzydatnych danych zawarte są cenne informacje, które można w odpowiedni sposób wykorzystać w różnych dziedzinach nauki i biznesu [Internet a].

Obecnie obserwuje się przejście od pojęcia przetwarzania danych do pojęcia przetwarzania i zarządzania wiedzą [Nycz, Smok 2004]. Przetwarzanie i zarządzanie wiedzą odnosi się do sposobu zdobywania, gromadzenia, uaktualniania, dzielenia się i operowania zasobami informacyjnymi, a przede wszystkim do przekształcania surowych danych w informację o charakterze strategicznym.

Dynamiczny rozwój technologii oraz technik informatycznych, a zwłaszcza sztucznej inteligencji, umożliwia przetwarzanie danych w wiedzę [Kiełtyka, Kulej 2002]. Technologie oraz techniki z zakresu sztucznej inteligencji ułatwiają odkrywanie wiedzy w bazach danych, zrozumienie struktury danych, konstruowanie wyjaśniających teorii, umożliwiają też modelowanie wiedzy oraz rozwiązywanie problemów niealgorytmizowanych na bazie symbolicznej reprezentacji wiedzy [Russel, Norvig 2002]. Na rynku teleinformatycznym dostępnych jest wiele programów, stanowiących oprogramowanie narzędziowe z zakresu sztucznej inteligencji. Jednym z nich jest zintegrowany pakiet SPHINX firmy Aitech. W jego skład wchodzą następujące moduły:

- system PC-Shell – szkieletowy system ekspertowy,
- system Neuronix – symulator sieci neuronowej,
- system CAKE – system komputerowego wspomagania inżynierii wiedzy,
- system HybRex – system do budowy inteligentnych aplikacji SWD (system wspomagania decyzji) i analizy danych,

- system Predyktor – system prognostyczny,
- system DeTreex – indukcyjny system pozyskiwania wiedzy.

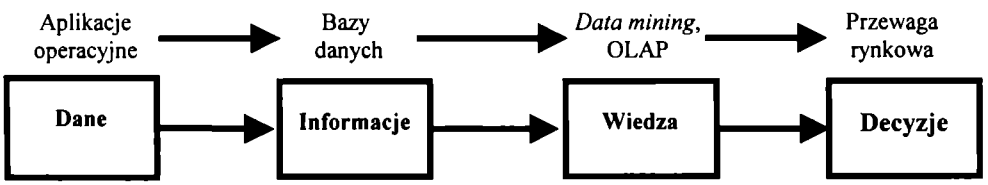
W niniejszym opracowaniu główną uwagę skoncentrowano na jednym z etapów przetwarzania danych w wiedzę – eksploracji danych. Podstawowym celem jest wskazanie uniwersalnych technik eksploracji danych i ich najważniejszych własności oraz przedstawienie wyników badań odkrywania struktury zbioru danych przy ich użyciu na przykładzie analizy problemu dotyczącego wdrożenia telepracy. W pracy wykorzystano systemy DeTreex i Neuronix, będące częścią pakietu Sphinx.

2. Proces przetwarzania danych w wiedzę

Jednym ze sposobów pozyskiwania wiedzy w przedsiębiorstwie jest analiza danych oparta na danych zgromadzonych w systemach baz danych. Do odkrywania wiedzy służy technologia eksploracji danych (*data mining* – DM), która umożliwia wygenerowanie wiedzy [Nycz, Smok 2004].

Dane to surowe informacje, zbudowane z liczb [Kiełtyka 2002]. Odpowiednio sklasyfikowane dane są informacjami wspomagającymi działalność przedsiębiorstwa. Informacja jest takim rodzajem zasobów przedsiębiorstwa, który pozwala na zwiększenie wiedzy o nim i jego otoczeniu [Kiełtyka 2002]. Odpowiednio przetworzona informacja pozwala na realizację funkcji zarządzania.

Nowoczesna technologia oraz techniki informatyczne umożliwiają przetwarzanie informacji w wiedzę. Wiedza to uporządkowany, posiadający swą strukturę wewnętrzną, spójny zbiór informacji, dotyczący określonej dziedziny, zagadnienia lub obserwowanego wycinka rzeczywistości [Nycz, Smok 2004]. Proces przetwarzania danych w wiedzę przedstawiono na rys. 1.



Rys. 1. Wiedza a procesy decyzyjne

Źródło: opracowanie własne na podstawie [Nycz, Smok 2004, s. 20].

3. Eksploracja danych

Jedną z technologii umożliwiającą analizę danych jest eksploracja danych, czyli *data mining* (DM). W DM wykorzystywane są metody statystyczne oraz metody sztucznej inteligencji, umożliwiające odkrywanie jeszcze nieznanych zależności pomiędzy danymi zgromadzonymi w zbiorach danych. W polskiej litera-

turze technologia DM jest określana terminami: eksploracja danych, odkrywanie wiedzy w bazach danych, drążenie danych, zgłębianie danych. Są to procesy analizy metodami zautomatyzowanymi dużych ilości danych ukrytych w strukturach baz danych w celu znalezienia istotnych wzorców, reguł, zależności, regularności, prawidłowości, powiązań czy anomalii [Nycz, Smok 2004]. Do podstawowych technik *data mining* zalicza się m.in. drzewa decyzyjne i sieci neuronowe [Sokołowski 2002].

Podstawowe własności drzew decyzyjnych. Drzewo decyzyjne definiuje się jako strukturę drzewiastą, której każdy węzeł odpowiada przeprowadzeniu pewnego testu na wartości jednego atrybutu, każdy zaś liść zawiera decyzję o klasyfikacji przykładu. Z poszczególnych węzłów wychodzi tyle gałęzi, ile jest możliwych wyników testu odpowiadających tym węzłom. Każda z tych gałęzi prowadzi do poddrzewa (węzła) służącego do klasyfikacji tych obiektów, dla których ten test ma określony wynik [Michalik 2000]. Dużą zaletą stosowania drzew decyzyjnych jest czytelność klasyfikowanych obiektów/zmiennych.

Wyróżniamy drzewa klasyfikacyjne i regresyjne. Jeżeli analizowana zmienna ma charakter ilościowy, to drzewo zbudowane w celu wyjaśnienia jej kształtowania się nazywane jest drzewem regresyjnym. Jeżeli natomiast zmienna objaśniana jest zmienną jakościową – określającą najczęściej przynależność do konkretnej klasy obiektów – to mamy do czynienia z drzewem klasyfikacyjnym [Sokołowski 2002].

Możliwość generowania drzew decyzyjnych zapewnia nam system DeTreex. System DeTreex jest narzędziem służącym do wspomagania procesu pozyskiwania wiedzy. Dzięki zastosowanej indukcyjnej metodzie tzw. uczenia maszynowego możliwe jest budowanie drzew decyzyjnych i zapis tych drzew w postaci reguł, które następnie mogą być bezpośrednio użyte do budowy baz wiedzy systemu ekspertowego PC-Shell. W ten sposób możemy utworzyć hybrydowy system wspomagania decyzji.

System DeTreex został opracowany z zastosowaniem metody indukcji drzew decyzyjnych. Indukcji dokonuje się na podstawie zgromadzonych wcześniej danych historycznych (ilościowych i jakościowych wartości atrybutów opisujących dany problem). System nie posiada żadnego ograniczenia liczby atrybutów i ich wartości oraz liczby rekordów w bazie danych, z jakich zostanie pozyskana wiedza.

DeTreex może zostać zastosowany do tworzenia baz wiedzy w tak różnych dziedzinach, jak np. ekonomia, finanse i bankowość, technika czy medycyna. Ogólnie system DeTreex może być stosowany wszędzie tam, gdzie pojawia się problem:

- podejmowania decyzji (klasyfikacji);
- szybkiego pozyskania reguł decyzyjnych z baz danych;
- szybkiej weryfikacji pozyskanych reguł.

Tabela 1. Zbiór danych wejściowych

Miejsce pracy	Czas pracy	Forma komunikacji	Relacja prawna	Decyzja
Dom, mieszkanie	synchroniczny	Poziom niski	umowa o pracę	nie
Telecentrum	synchroniczny	Poziom niski	umowa o pracę	nie dotyczy
Wszędzie	synchroniczny	Poziom niski	umowa o pracę	nie
Dom, mieszkanie	synchroniczny	Poziom wysoki	umowa o pracę	telepraca domowa
Telecentrum	synchroniczny	Poziom wysoki	umowa o pracę	telepraca w telecentrum
Wszędzie	synchroniczny	Poziom wysoki	umowa o pracę	nie
Dom, mieszkanie	synchroniczny	Poziom niski	umowa cywilnoprawna	nie
Telecentrum	synchroniczny	Poziom niski	umowa cywilnoprawna	nie dotyczy
Wszędzie	synchroniczny	Poziom niski	umowa cywilnoprawna	nie
Dom, mieszkanie	synchroniczny	Poziom wysoki	umowa cywilnoprawna	nie
Telecentrum	synchroniczny	Poziom wysoki	umowa cywilnoprawna	telepraca w telecentrum
Wszędzie	synchroniczny	Poziom wysoki	umowa cywilnoprawna	nie dotyczy
Dom, mieszkanie	asynchroniczny	Poziom niski	umowa o pracę	telepraca domowa
Telecentrum	asynchroniczny	Poziom niski	umowa o pracę	nie dotyczy
Wszędzie	asynchroniczny	Poziom niski	umowa o pracę	telepraca mobilna
Dom, mieszkanie	asynchroniczny	Poziom wysoki	umowa o pracę	telepraca domowa
Telecentrum	asynchroniczny	Poziom wysoki	umowa o pracę	telepraca w telecentrum
Wszędzie	asynchroniczny	Poziom wysoki	umowa o pracę	nie dotyczy
Dom, mieszkanie	asynchroniczny	Poziom niski	umowa cywilnoprawna	telepraca domowa
Telecentrum	asynchroniczny	Poziom niski	umowa cywilnoprawna	nie dotyczy
Wszędzie	asynchroniczny	Poziom niski	umowa cywilnoprawna	telepraca mobilna
Dom, mieszkanie	asynchroniczny	Poziom wysoki	umowa cywilnoprawna	telepraca domowa
Telecentrum	asynchroniczny	Poziom wysoki	umowa cywilnoprawna	telepraca w telecentrum
Wszędzie	asynchroniczny	Poziom wysoki	umowa cywilnoprawna	nie dotyczy

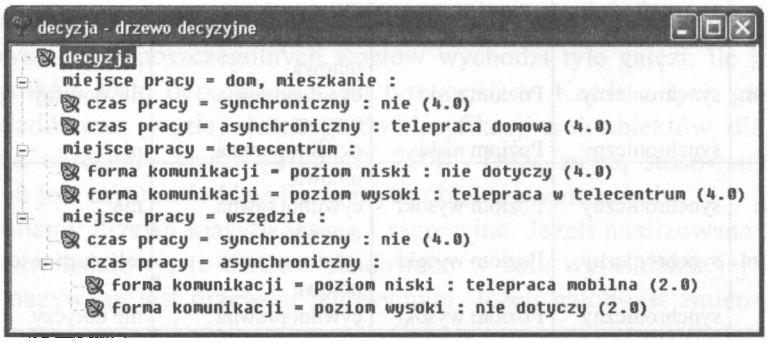
Źródło: opracowanie własne.

Przykład zastosowania drzew decyzyjnych. W ramach badań własnych dotyczących wdrażania telepracy w polskich organizacjach gospodarczych wyróżniono kryteria klasyfikacji tej formy zatrudnienia. Do kryteriów tych zaliczono:

miejsce pracy, czas pracy, formę komunikacji i relację prawną między pracownikiem a pracodawcą. W tabeli 1 przedstawiono zbiór przykładów uczących opracowany na podstawie w/w kryteriów.

Na podstawie zbioru uczącego przedstawionego w tab. 1 zostało wygenerowane drzewo decyzyjne. Do wygenerowania drzewa decyzyjnego wykorzystano system DeTreex.

Podczas budowy drzewa decyzyjnego przeprowadzany jest proces uogólnienia wiedzy na podstawie wcześniej zgromadzonych danych historycznych (ilościowych i/lub jakościowych). Drzewo w postaci tekstowej miało postać jak na rys. 2.



Rys. 2. Drzewo decyzyjne w postaci tekstowej

Źródło: opracowanie własne.

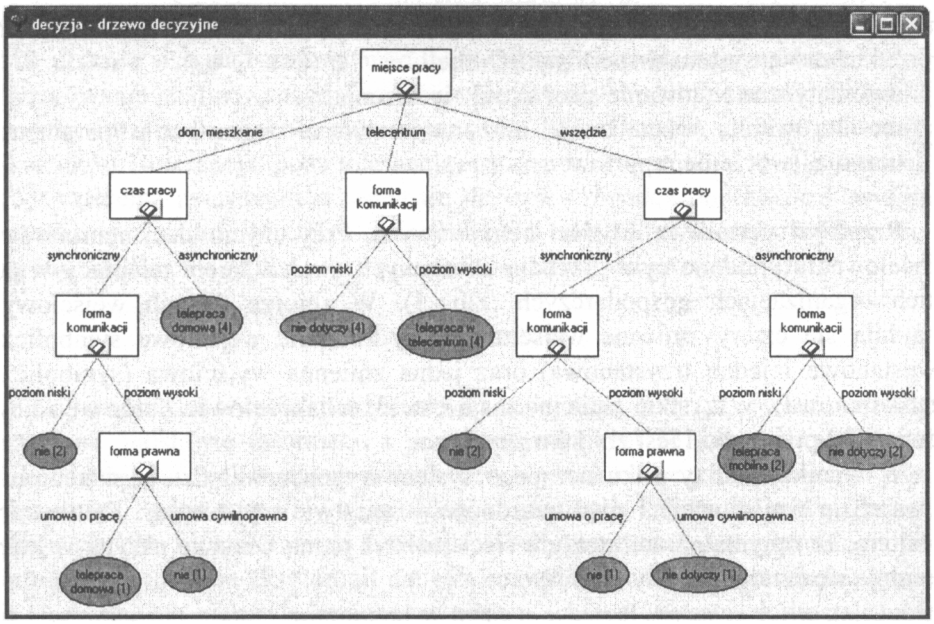
W ramach budowy drzewa decyzyjnego przyjęto następujące parametry:

- przycięcie drzewa decyzyjnego – nie przycinaj drzewa;
- minimalna liczba przykładów tworzących liść drzewa – 2.

Opcja „przycięcie drzewa decyzyjnego" umożliwia zmniejszenie rozmiarów drzewa po wygenerowaniu tego drzewa, a tym samym zwiększenie stopnia uogólnienia wiedzy reprezentowanej przez drzewo. W rozważanym przykładzie dokonano wyboru „nie przycinaj drzewa", co oznacza, że zbudowane drzewo nie zostanie zmniejszone i będzie zachowany jego pierwotny kształt. Parametr „minimalna liczba przykładów tworzących liść drzewa" decyduje o rozmiarze drzewa. Im wartość tego parametru jest większa, tym wygenerowane drzewo jest mniejszych rozmiarów (w ten sposób osiąga się duży stopień uogólnienia wiedzy). Dla porównania na rys. 3 pokazano postać graficzną drzewa, dla którego przyjęto następujące parametry:

- przycięcie drzewa decyzyjnego – nie przycinaj drzewa;
- minimalna liczba przykładów tworzących liść drzewa – 1.

Drzewo w postaci graficznej, które zostało wygenerowane przy innych parametrach, jest drzewem, które w mniejszym stopniu uogólnia wiedzę na podstawie zgromadzonych danych.



Rys. 3. Drzewo decyzyjne w postaci graficznej

Źródło: opracowanie własne.

Podstawowe własności sieci neuronowych. Mianem sieci neuronowych (SN) określa się symulatory (programowe lub sprzętowe) modeli matematycznych, realizujące równoległe przetwarzanie informacji, składające się z wielu wzajemnie połączonych neuronów (naśladują działanie biologicznych struktur mózgowych). Podstawową cechą różniącą SN od programów realizujących algorytmiczne przetwarzanie informacji jest zdolność uogólniania wiedzy dla nowych danych nieznanych wcześniej, czyli nie prezentowanych w trakcie nauki. Określa się to także jako zdolność SN do aproksymacji wartości funkcji wielu zmiennych w przeciwieństwie do interpolacji możliwej do otrzymania przy przetwarzaniu algorytmicznym [Kosiński 2002; Osowski 2000].

Programem umożliwiającym budowanie układów symulujących działanie sieci neuronowej jest system Neuronix. System Neuronix umożliwia wszechstronną analizę danych poprzez tworzenie modeli różnorodnych zjawisk występujących m.in. w ekonomii i technice. Zastosowanie sieci neuronowej pozwala na automatyczne utworzenie modelu bez konieczności szczegółowej znajomości modelowanego zjawiska [Witkowska 2000].

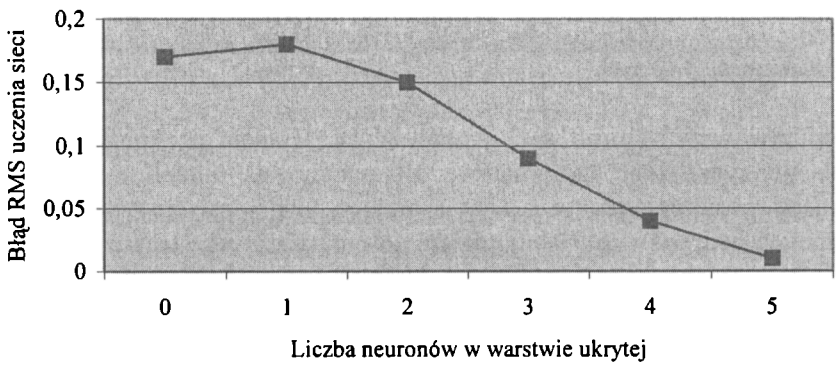
Główne funkcje systemu Neuronix to:

- wspomaganie procesu projektowania sieci, począwszy od gromadzenia próbek, poprzez generację plików uczących i testowych, uczenie i testowanie sieci, do jej uruchamiania dla wskazanych wartości wejściowych w celu wnioskowania,

- budowa aplikacji hybrydowych wykorzystujących sieć neuronową oraz szkieletowy system ekspertowy PC-Shell,
- automatyczne testowanie sieci w trakcie uczenia,
- monitorowanie, wizualizacja oraz automatyczne zapisywanie parametrów uczenia, tworzenie raportów.

Przykład zastosowania sieci neuronowych. Przy użyciu sieci neuronowych modelowaniu poddano cykl procedur związanych z wdrażaniem telepracy w polskich organizacjach gospodarczych (tab. 1). W zbiorze danych wejściowych znajdują się cztery zmienne wejściowe (trzy zmienne wejściowe symboliczne dwustanowe i jedna trzystanowa) oraz jedna zmienna wyjściowa (symboliczna pięciostanowa). Wszystkie zmienne mają charakter jakościowy. Zbiór wejściowy zawiera 24 przypadki i jest zbiorem zupełnym.

W wyniku analizy pliku uczącego system wygenerował dziewięć neuronów w warstwie wejściowej i pięć neuronów w warstwie wyjściowej. Empirycznie ustalono, że optymalna architektura sieci zawiera jedną warstwę ukrytą, w której znajdują się cztery neurony (rys. 4).



Rys. 4. Wykres błędu RMS uczenia sieci w zależności od liczby neuronów znajdujących się w warstwie ukrytej

Źródło: opracowanie własne.

Przyjęto wartości domyślne współczynników uczenia (generowane przez program). Załączono opcję mieszania wzorców. Każdy przypadek jest niezależny, a budowany model ma charakter przyczynowo-skutkowy.

W projekcie ustalono warunek zatrzymania jako: RMS uczenia < 'epsilon'. Przyjęto wartość progu liczbowego epsilon równą 0,01. Jeżeli błąd RMS uczenia dla badanej architektury sieci nie osiągnął wartości mniejszej od zadanej, proces uczenia sieci kończono w momencie, gdy nie obserwowano już żadnej poprawy jakości sieci.

Dla modelu z jedną warstwą ukrytą, w której znajduje się pięć neuronów, błąd RMS uczenia osiągnął zadaną wartość progową, co wskazuje na bardzo dobre odwzorowanie analizowanego zjawiska – zbudowany model poprawnie klasyfikuje wszystkie przypadki (jakość sieci wynosi 1). Na podstawie analizy pozostałych wskaźników (tolerancja, poza tolerancją) ustalono, iż optymalna topologia sieci to sieć z czterema neuronami w warstwie ukrytej – również poprawnie klasyfikuje wszystkie przypadki. Model z jedną warstwą ukrytą, w której znajdują się 3 neurony, błędnie klasyfikuje dwa przypadki – jakość sieci równa 0,92.

4. Zakończenie

W przedsiębiorstwach menedżerowie podejmują decyzje, zwłaszcza strategiczne, na podstawie informacji i wiedzy wynikającej ze szczegółowych analiz wielu czynników obejmujących makrootoczenie, mikrootoczenie oraz samo przedsiębiorstwo. Podczas analizy zbioru danych pewnych zależności nie widać albo wydają się one niemożliwe do istnienia. Wtedy przydatna jest technologia *data mining*. Zastosowanie eksploracji danych pozwala na odkrywanie wiedzy uprzednio nieznanej lub nieuświadomionej w postaci schematów, związków, zależności, anomalii czy struktur.

W opracowaniu pokazano ogólne możliwości oraz właściwości drzew decyzyjnych i sieci neuronowych na podstawie funkcjonowania systemu DeTreex i Neuronix. Bazowano na pakiecie SPHINX z uwagi na dostępność tego programu (relatywnie niska cena), jego wszechstronność (dobrze zintegrowane moduły z różnych dziedzin) i jednocześnie prostotę (ważne szczególnie w przypadku małych firm, które cechuje mała dostępność wysoko wykwalifikowanej kadry). Zastosowanie drzew decyzyjnych wskazuje na łatwość i dużą czytelność tej techniki DM. Wizualizacja danych sprawia, że potencjalni decydenci dostają gotową wiedzę o zależnościach pomiędzy klasyfikowanymi obiektami. Drzewa decyzyjne pozwalają na wyselekcjonowanie ze zbioru zmiennych objaśniających tych zmiennych, które mają decydujący wpływ na interesującą nas zmienną zależną (w rozpatrywanym przypadku była to decyzja o możliwości wdrożenia telepracy i jej rodzaju) [Internet b].

Sieci neuronowe również odwzorowały istniejącą strukturę pomiędzy danymi, ale drzewa decyzyjne opisują rozwiązanie danego zadania klasyfikacyjnego w sposób bardziej przejrzysty.

Drzewa decyzyjne oraz sieci neuronowe mogą stanowić niezależne systemy decyzyjne, jak też mogą być wykorzystywane do automatyzacji procesu budowania baz wiedzy w systemach hybrydowych.

Literatura

Internet a, http://www.pckurier.pl/archiwum/art0.asp?ID=5741.
Internet b, http://sszymanski.strony.wi.ps.pl/kdd.html.
Kiełtyka L., *Komunikacja w zarządzaniu. Techniki, narzędzia i formy przekazu informacji*, Placet, Warszawa 2002.
Kiełtyka L., *Zarządzanie danymi w nowoczesnym przedsiębiorstwie*, [w:] *Nowoczesne zarządzanie przedsiębiorstwem* (część 2), red. J. Stankiewicz, Redakcja Wydawnictw Nauk Ścisłych i Ekonomicznych, Zielona Góra 2001.
Kiełtyka L., Kulej E., *Systemy ekspertowe i Business Intelligence – wykorzystanie zasobów danych przedsiębiorstwa do wspomagania decyzji biznesowych*, X Konferencja Naukowo-Techniczna, Produkcja i Zarządzanie w Hutnictwie, Ustroń-Jaszowiec 2002.
Kosiński R.A., *Sztuczne sieci neuronowe*, WNT, Warszawa 2002.
Michalik K., *DeTreex 3.0. dla Windows 9x/NT/2000. Indukcyjny system pozyskiwania wiedzy, Podręcznik użytkownika*, Katowice 2000.
Osowski S., *Sieci neuronowe do przetwarzania informacji*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2000.
Nycz M., Smok B., *Wykorzystanie narzędzi Data Mining do odkrywania wiedzy wspomagającej decydenta*, [w:] *Komputerowo zintegrowane zarządzanie*, red. R. Knosali, WNT, Warszawa 2004.
Russel S., Norvig P., *Artificial Intelligence. A Modern Approach*, Prentice Hall Englewood Cliffs 2002.
Sokołowski A., *Metody stosowane w Data Mining*, [w:] *Data Mining – Metody i przykłady*, StatSoft, Kraków 2002.
Witkowska D., *Sztuczne sieci neuronowe i metody statystyczne*, C.H. Beck, Warszawa 2002.

DATA MINING WITH THE USE OF SPHINX PACKET

Summary

In this paper attention was paid to the process of processing data into knowledge. One of the stage of this process is Data Mining. In this paper the main property of chosen techniques of Data Mining (for example decision trees and neural networks) was shown. The idea of using these techniques was presented on the grounds of analysis of the problem of telework carrying out. The program Sphinx was used.

Dr inż. Klaudia Smoląg jest adiunktem w Katedrze Informatycznych Systemów Zarządzania na Wydziale Zarządzania Politechniki Częstochowskiej
e-mail: klaudia@zim.pcz.czest.pl

Mgr inż. Waldemar Jędrzejczyk jest asystentem w Katedrze Informatycznych Systemów Zarządzania na Wydziale Zarządzania Politechniki Częstochowskiej
e-mail: waldekj@zim.pcz.czest.pl