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

Małgorzata Nycz, Barbara Smok

METODY ANALIZY I EKSPLORACJI DANYCH

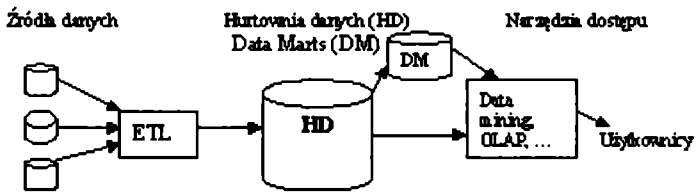
1. Wprowadzenie

W rzeczywistości gospodarczej opartej na wiedzy zdolność przedsiębiorstwa do tworzenia wartości jest zależna od szybkiej reakcji na zmiany warunków rynkowych, jak również od wykorzystania aktywów rzeczowych i niematerialnych. Systemy zarządzania bazami danych gromadzą dane transakcyjne, wspomagając codzienną działalność przedsiębiorstwa. Przedsiębiorstwa posiadają olbrzymie zbiory danych w bazach danych i innych dokumentach, które stanowią skarbnicę informacji na temat procesów zachodzących w przedsiębiorstwie. Wykorzystanie tych informacji może mieć wpływ na wzrost efektywności procesów gospodarczych oraz umożliwiać przedsiębiorstwu przeprowadzenie badań trendów zachodzących wewnątrz i na zewnątrz firmy, jak również badanie wzajemnego ich wpływu. Informacje te są ukryte wśród ogromu danych. Podejmując w przedsiębiorstwie różnego rodzaju decyzje, wykorzystujemy naszą wiedzę nabytą w czasie wcześniejszych doświadczeń. Systematyczne badanie kondycji przedsiębiorstwa, wpływu otoczenia, wykorzystanie nadarzających się okazji, przeciwdziałanie zagrożeniom wewnętrznym i zewnętrznym nie jest dzisiaj możliwe bez systemu informatycznego oraz narzędzi analitycznych. Cele nowoczesnych inicjatyw strategicznych, takich jak CRM, SCM i *e-business*, mogą być osiągnięte wówczas, gdy decydenci będą wspierani różnymi analizami. Hurtownia danych zapewnia mechanizmy efektywnego przetwarzania oraz dostarcza technologii informatycznych wspomagających zaawansowane analizy i eksplorację danych.

2. Hurtownie danych

Hurtownia danych (HD) jest zorientowaną tematycznie, integralną, uporządkowaną w czasie, nieulotną kolekcją danych, wspierającą proces podejmowania

decyzji przez kierownictwo [Inmon 2002, s. 17]. Jest niezależna od źródeł danych, dostępna dla użytkowników biznesowych, zintegrowana zgodnie z modelem korporacyjnym, gromadząca dane w czasie (zarówno historyczne, jak i aktualne) oraz łatwo dostępna dla różnych użytkowników, nie tylko dla informatyków [Inmon, Welch, Glassey 19970]. Przykład hurtowni danych zaprezentowano rys. 1.



Rys.1. Hurtownia danych

Źródło: opracowanie własne.

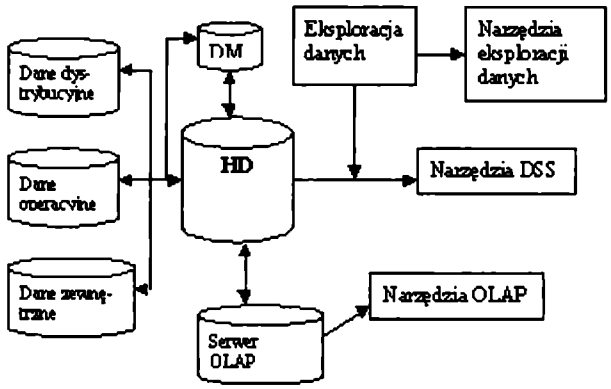
Hurtownia danych zapewnia natychmiastowy dostęp do informacji, od której zależy podejmowanie decyzji, należy ją zatem traktować jako specyficzną bazę danych, która przechowuje i udostępnia dane o przebiegu procesów w firmie – z myślą o obsłudze procesów decyzyjnych [Nycz, Smok 2000, s. 9; Smok 2003 s. 415]. Narzędzia dostępu do danych to narzędzia do raportowania i zapytań *ad hoc*, narzędzia OLAP (*on-line analytical processing*), narzędzia eksploracyjne umożliwiające różnego rodzaju analizy, które mogą być wykorzystane przez użytkowników biznesu.

HD jest nowym podejściem do zbierania informacji i tworzenia raportów na podstawie danych zgromadzonych w korporacyjnych systemach komputerowych. Różni się od tradycyjnych systemów raportowania tym, że umożliwia użytkownikowi bezpośredni dostęp do danych korporacyjnych za pomocą narzędzi do zadawania zapytań i tworzenia raportów oraz wymaga stworzenia oddzielnej bazy danych przeznaczonej do wspierania procesów podejmowania decyzji, uzyskanej z połączenia danych wielu systemów operacyjnych. Hurtownia danych pozwala również użytkownikowi na przeglądanie ogromnych zasobów danych korporacyjnych w sposób interaktywny w celu określenia tendencji rozwoju, rozwiązania problemów, oceny szans rynkowych. Wynikiem jest możliwość szybkiego podejmowania lepiej uzasadnionych decyzji.

Składnice danych (*data marts*), zwane też hurtowniami tematycznymi lub minihurtowniami, najczęściej posługują się tabelami zdenormalizowanymi i hierarchicznymi strukturami gwiaździstymi, zapewniającymi szybkie uzyskiwanie odpowiedzi na skomplikowane zapytania analityczne, wymagające dostępu do danych i ich agregatów (sum cząstkowych). Zawierają one dane archiwalne mocno zagregowane w wielu wymiarach, co pozwala na szybkie uzyskiwanie odpowiedzi

na wiele zapytań standardowych. Tabele faktów umożliwiają dostęp do szczegółowych danych transakcyjnych, „drążenie w głąb" i uzyskiwanie odpowiedzi na coraz bardziej szczegółowe pytania. HD ma na celu zapewnienie źródła danych – ujednoliconych, oczyszczonych z błędów i niejednoznaczności, przeniesionych z systemów transakcyjnych. Znajdujące się w HD dane pochodzące z heterogenicznych źródeł i będące w różnych formatach są konwertowane do spójnego, zunifikowanego formatu oraz zorganizowane w sposób pozwalający na łatwy dostęp do informacji wspomagających procesy decyzyjne. Utrzymanie hurtowni danych jest procesem zaspokajania potrzeb przedsiębiorstwa w procesach podejmowania decyzji poprzez udostępnianie wiedzy wygenerowanej w różnych analizach. Hurtownia danych jest to zintegrowany bank danych, zorientowany na analizowanie i wyodrębnianie informacji pochodzących z różnych źródeł. Informacje te zostają w określony sposób konwertowane do jednolitego formatu, a następnie zorganizowane w sposób pozwalający na łatwy do nich dostęp, co z kolei wspomaga procesy decyzyjne.

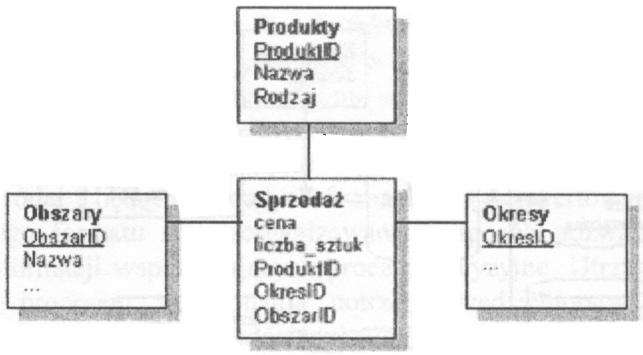
HD jest jedną z technologii umożliwiających integrację danych w różnych systemach informacyjnych funkcjonujących w przedsiębiorstwie. Jednym z celów integracji jest umożliwienie różnych wielowymiarowych analiz czy eksploracji na potrzeby wspomagania procesów podejmowania decyzji [Inmon 2002, s. 20-40]. Tradycyjny sposób korzystania z baz danych to najczęściej realizacja zapytań poprzez aplikacje lub raporty, a więc model OLTP (*on-line transaction processing*) – przetwarzanie transakcji w trybie *on-line*. Model ten świetnie nadaje się do bieżącej działalności przedsiębiorstwa, lecz nie wspomaga procesów analizy danych. Do analizy wielowymiarowej wykorzystuje się model OLAP, którego głównym zadaniem jest efektywne dostarczenie strategicznej informacji i zaprezentowanie jej zgodnie z ludzkimi schematami poznawczymi. OLAP cechuje funkcjonalność analityczna i zdolność do przedstawiania danych w formacie wielowymiarowym. Jest on dobrym rozwiązaniem do obsługi istotnych decyzji biznesowych. Przy nadmiarowości danych zgromadzonych w obecnych systemach informatycznych często utrudnione jest dotarcie do informacji niezbędnej do udzielenia odpowiedzi na pytania nurtujące osoby podejmujące decyzje, które często są kluczem do utrzymania się firmy na dynamicznym rynku. Narzędzia typu OLAP pozwalają użytkownikom na analizę danych w celu wykrycia trendów, odnalezienia wyjątków, uzyskania szczegółów czy przeglądania podsumowań. OLAP pomaga m.in. w analizie działalności przedsiębiorstwa, analizie trendów w zarządzaniu i opracowywaniu strategii przedsiębiorstwa. Technologia ta umożliwia użytkownikom HD tworzenie *ad hoc* interakcyjnych zapytań analitycznych pokazujących przekroje (np. statystyczne) danych, np. szeregi czasowe lub trendy. Powiązanie pomiędzy HD a narzędziami OLAP zaprezentowano na rys. 2.



Rys. 2. Hurtownia danych i OLAP
Źródło: opracowanie własne na podstawie [Bębel, Morzy 2000].

W modelu OLAP dane są postrzegane przez użytkowników w postaci wielowymiarowej perspektywy (tzw. kostki OLAP). Obiektem analizy jest tutaj zbiór miar numerycznych. Miara jest podstawowym pojęciem schematu pojęciowego systemu OLAP. Ma charakter liczbowy (np. ilość sprzedanych produktów, średnia ocen studentów, średnie zarobki). Z każdą miarą jest związany zbiór wymiarów, od których zależy wartość danej miary (np. ilość sprzedanych produktów w zależności od produktu, czasu sprzedaży czy miejsca sprzedaży). Wymiarami mogą być produkt, lokalizacja, czas. Z każdym wymiarem jest związany zbiór atrybutów. Atrybuty opisujące pojedynczy wymiar tworzą hierarchię nazywaną hierarchią wymiaru, która umożliwia definiowanie różnych poziomów agregacji danych (zasadniczy cel budowy systemu OLAP). Hierarchia wymiaru może być reprezentowana w HD *implicite* – w pojedynczej tablicy wymiaru lub *explicite* – w postaci zbioru powiązanych tablic reprezentujących jeden wymiar. Fakty stanowią najistotniejszy obszar danych w HD, gdyż są podstawą do wszelkich analiz. Fakt opisuje pojedyncze zdarzenie, o którym informację chcemy przechować w HD. Jest on daną ilościową (numeryczną) reprezentującą jednostkę aktywności biznesowej przedsiębiorstwa (np. średnia ocena studenta, zysk, wartość produktu krajowego itp.).

Każdy wymiar jest, jak powiedzieliśmy, opisany zbiorem atrybutów. Mogą one mieć bardzo dużą objętość, nawet kilku terabajtów (w zależności od tego, jak dużo danych historycznych jest potrzebnych do analizy). Mogą być fizycznie podzielone (partycjonowane) na mniejsze tablice, a logicznie reprezentowane jako jedna. Pola w tabelach zawierających dane dotyczące faktów są w dużym stopniu indeksowane, tak aby przyspieszyć realizację zapytań typu *ad hoc*. Tablica faktów odzwierciedla w HD aspekt dynamiczny świata rzeczywistego, natomiast tablice wymiarów reprezentują aspekt statyczny.



Rys. 3. Przykład gwiazdy

Źródło: opracowanie własne na podstawie [Morzy 2002].

Architektura hurtowni powinna zapewniać elastyczność i skalowalność, dzięki czemu może wykorzystać wiele różnych typów danych i różnych metod do nich dostępu. Projekt hurtowni danych powinien uwzględniać różne cele biznesowe w sensowny sposób. HD musi odzwierciedlać model działalności końcowego użytkownika. Hurtownie danych wspierają różnego rodzaju aplikacje analityczne. Jako najbardziej popularne zagadnienia można wymienić: zarządzanie relacjami z klientami, analizę sprzedaży, analizę zyskowności, kontrolę zadłużeń, monitorowanie kosztów, segmentację rynku, analizę promocji i kampanii, wspomaganie raportowania finansowego, ocenę ryzyka oraz wykrywanie nadużyć.

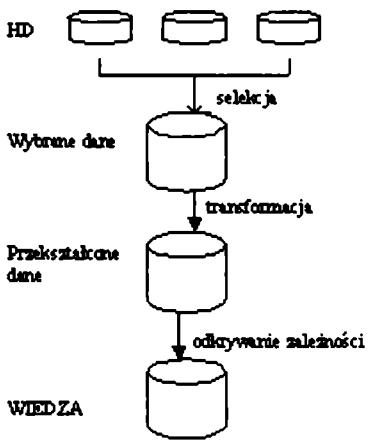
Aplikacje analityczne wymagają integracji danych, złożonej analizy i eksploracji danych.

4. Eksploracja danych w hurtowni danych

W obszarze zastosowań obserwujemy szybki rozwój systemów informacyjnych, zwłaszcza za sprawą Internetu. Przedsiębiorstwa gromadzą informacje długotrwałe, w związku z czym istniejące zasoby informacyjne zawarte w różnorakich bazach danych są niezwykle duże i stale rosną. Wykorzystanie tych danych do różnego rodzaju analiz jest trudnym, czasochłonnym i kosztownym przedsięwzięciem; jednak ich niewykorzystanie pozbawi nas wiedzy, która jest w nich ukryta i która może być użyteczna.

Mianem **eksploracji danych** określa się poszukiwanie wiedzy ukrytej głęboko, gdzieś w gigabajtowych bazach. Wiedza to coś więcej niż informacja, to pewna struktura, a więc specyficzne korelacje, prawidłowości statystyczne lub inne zależności, które dają się wyrazić w języku matematyki lub w dowolnym języku naturalnym. Czasami trudno do nich dotrzeć, a niekiedy nie podejrzewa się nawet ich istnienia [Han 1999, s. 20-30]. Dane te mogą mieć realną wartość liczoną w tysiącach złotych, np. jeśli dotyczą ważnych dla jakiegoś sektora zachowań

rynkowych. Ich znalezienie może oznaczać umiejętność przewidywania przyszłości, a tym samym mogą dać znaczącą przewagę nad konkurencją [Berry, Linoff 1997, s. 10-20].



Rys. 4. Proces eksploracji danych z hurtowni danych

Źródło: opracowanie własne.

Dane w HD zwykle wykorzystywane są do różnego rodzaju analiz, jak np. złożona analiza wielowymiarowa czy też eksploracja danych, w celu wygenerowania wiedzy potrzebnej do wspomagania decyzji. Eksploracja danych (drążenie danych, zgłębianie danych, *data mining*) jest techniką analizy specyficzną dla HD [Nycz, Smok 2004, s. 200]. Dostęp do niej zapewniają: specjalna struktura danych oraz interaktywne narzędzia OLAP [Oracle 2002, s. 50]. Przez eksplorację najczęściej rozumie się proces automatycznego odkrywania dotychczas nieznanej, lecz znaczącej i użytecznej wiedzy w dużych bazach danych.

W procesie odkrywania wiedzy pochodzącej z różnych źródeł określa się sposoby jej zdobywania, dzielenia się i operowania tymi zasobami. Wymaga to integracji wiedzy pochodzącej m.in. z baz danych, hurtowni danych, systemów uczących się czy statystyki. W sytuacji, gdy mamy do czynienia z wielowymiarowymi zbiorami danych, niezbędne jest zastosowanie metod umożliwiających odkrywanie istotnej wiedzy w postaci schematów, związków, zależności czy struktur. Odkrywanie wiedzy to pozyskiwanie użytecznej wiedzy z wykorzystaniem danych zgromadzonych w różnych w bazach danych [Kantardzic 2002, s. 21-25]. Jest to proces, w skład którego wchodzi wiele etapów (rys. 4), jak: gromadzenie danych, czyszczenie (m.in. obsługa błędnych lub brakujących danych), integracja (łączenie danych pochodzących z różnych źródeł), selekcja (wybranie danych istotnych ze względu na analizowany problem), transformacja (nadanie odpowiedniej reprezentacji wyselekcjonowanym danym), „drążenie" (wykorzy-

stanie 'inteligentnych' metod przetwarzania danych celem uzyskania m.in. reguł, schematów, zależności), weryfikacja (interpretacja wyników) oraz prezentacja wiedzy (zastosowanie technik wizualizacji i reprezentacji wiedzy użytkownikowi).

Odkryta w ten sposób wiedza przyjmuje różną postać, np. reguł, prawidłowości, tendencji, korelacji. Odkryte zależności nie muszą być poprawne, dlatego też zachodzi potrzeba weryfikacji spójności, skuteczności i użyteczności zastosowanego rozwiązania. Taką weryfikację można przeprowadzić poprzez symulację komputerową – z wykorzystaniem wiedzy dziedzinowej – lub też w rzeczywistych warunkach [Byrski 2002, s. 113-122].

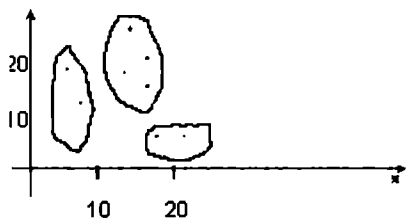
Istnieje wiele różnych algorytmów odkrywania wiedzy. Jedną z klasyfikacji algorytmów odkrywania wiedzy może być podział na [Chen 1996]: odkrywanie zależności (*mining association rules*), wielopoziomowe uogólnianie danych (*multi-level data generalization*), klasyfikacja (*data classification*), grupowanie (*clustering analysis*), odkrywanie podobieństw na podstawie wzorca (*pattern similarity search*), odkrywanie schematów ścieżek (*mining path traversal patterns*).

Odkrywanie zależności np. w bazie danych o transakcjach przeprowadzonych w sklepie będzie polegać na identyfikacji artykułów nabywanych łącznie (np. garnitur i koszula). Jeśli $A = \{a_1, a_2, ..., a_n\}$ będzie zbiorem elementów reprezentujących artykuły w sklepie, a $T = \{T_1, T_2, ..., T_n\}$ będzie zbiorem transakcji reprezentujących fakt zakupienia co najmniej dwóch artykułów, to oznacza, że $T_i \subset A$. Jeśli założymy, że $X \subset A$, wówczas o transakcji T_i możemy powiedzieć, że zawiera zbiór X wtedy i tylko wtedy, gdy $X \in T_i$. Zależność tę możemy przedstawić w formie implikacji $X \Rightarrow Y$, gdzie $X \subset A$, $Y \subset A$, a $X \cap Y = \emptyset$. O zależności $X \Rightarrow Y$ można powiedzieć, że posiada wiarygodność c (gdy c% transakcji ze zbioru T zawierających podzbiór X – zawiera również podzbiór Y – informujący nas o sile zależności) oraz wsparcie o wartości s (jeśli s% transakcji ze zbioru T zawiera podzbiór X lub Y – informujące nas o częstości pojawiania się zależności w bazie danych). Główne zadanie algorytmu *data mining* to znalezienie silnych zależności, które charakteryzuje duża wiarygodność i silne wsparcie, a więc zidentyfikowanie największych zbiorów elementów w bazie o wsparciu powyżej wyznaczonej granicy i wykorzystanie ich do wygenerowania poszukiwanych zależności.

Algorytmy służące do wielopoziomowego uogólniania danych oprócz cech odkrywania zależności mogą zawierać elementy ułatwiające przeprowadzanie analiz, takie jak np. [Han 1995]: generowanie zależności zbudowanych na różnych poziomach abstrakcji, definiowanie różnych minimalnych wartości wsparcia dla różnych poziomów hierarchii, warunkowe badanie zależności na niższym poziomie wówczas, gdy ta zależność posiada na wyższym poziomie odpowiednie wsparcie.

Klasyfikacja danych jest procesem, którego celem jest znalezienie wspólnych cech charakterystycznych wśród obiektów bazy danych i przyporządkowanie ich do odpowiednich klas (grup), które pozwolą odróżnić je od pozostałych klas obiektów. Wśród metod klasyfikacji możemy wyróżnić drzewa decyzyjne, których tworzenie odbywa się przez rekurencyjny podział zbioru na podzbiory aż do

uzyskania ich jednorodności ze względu na przynależność obiektów do klas. Klasyfikacja jest znana i wykorzystywana w sztucznej inteligencji i uczeniu maszynowym. Wykorzystuje się ją do weryfikacji kredytobiorców, podziału pacjentów. Klastrowanie często wiąże się z klasyfikacją, gdzie obiekty przypisane są do klas. W przypadku klastrowania jest to znajdowanie skończonego zbioru klas obiektów (klastrów) w bazie danych o zbliżonych cechach. Klastrowanie najczęściej wykorzystuje się do określania segmentów rynku na podstawie cech klientów. Na rysunku 5 przedstawiono trzy klasy obiektów: klasa1: $x < 10$, klasa 2: $10 < x < 20$ i $y > 10$, klasa 3: $y < 10$. Problem klastrowania danych czasami nazywamy taksonomią danych. Każdy obiekt w otaczającym nas świecie możemy przyporządkować do pewnej grupy elementów, które posiadają pewne cechy, np. produkty spożywcze możemy podzielić na pieczywo, nabiał i wędliny.

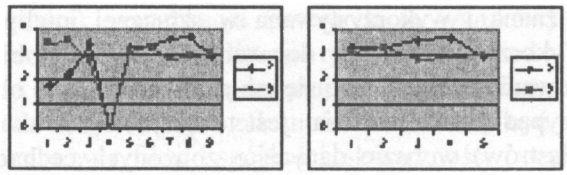


Rys. 5. Przykład klastrowania

Źródło: opracowanie własne.

Grupowanie pozwala na utworzenie grup zdarzeń lub podobnych do siebie obiektów ze względu na określone kryteria. Wykorzystując określony sposób pomiaru odległości (podobieństwa) obiektów w wielowymiarowej przestrzeni cech, można podzielić zbiór na podzbiory tak, aby zawierały obiekty najbardziej do siebie podobne. Możemy wykorzystać tutaj takie techniki, jak [Gatn 1998]:

- Metoda odkrywania podobieństw w przebiegach czasowych wykorzystywana jest do grupowania „podobnych" przebiegów czasowych (np. wykresów), które mogą być przesunięte w czasie, ale mają podobną charakterystykę. Metodę tę można wykorzystać m.in. do grupowania firm na giełdzie o podobnej dynamice wzrostu cen akcji czy surowców o podobnej charakterystyce sprzedaży (rys. 6).
- Metody eksploracji pozwalają również na wykrywanie zmian i odchyleń polegających na wyszukiwaniu różnic pomiędzy aktualnymi i oczekiwanymi wartościami danych. Przykładem może być wyszukiwanie anormalnych zachowań klientów kart kredytowych. Stosuje się je głównie do analizy danych wielowymiarowych. Przez odchylenie rozumie się różnicę pomiędzy aktualną a oczekiwaną wartością.



Rys. 6. Wykorzystanie szeregów czasowych do odkrywania wiedzy
Źródło: opracowanie własne.

- Metodę odkrywania podobieństw na bazie wzorców najczęściej wykorzystuje się do analizy szeregów czasowych, a więc zbiorów danych, w których jednym z atrybutów jest czas lub też inny atrybut zależny od czasu. Możemy mieć tutaj do czynienia z dwoma przypadkami [Chen, Han, Ju 1996]: zapytaniami związanymi z określonym wzorcowym obiektem, których celem jest znalezienie obiektów spełniających wcześniej zdefiniowane warunki dotyczące podobieństwa do wzorcowego obiektu, lub zapytaniami porównującymi wszystkie pary elementów ze sobą, których celem jest znalezienie par obiektów spełniających określony przez użytkownika warunek podobieństwa.

Wykorzystanie metod sztucznej inteligencji do odkrywania wiedzy z baz danych może przynieść zyski wykorzystującym je przedsiębiorstwom. *Data mining* jest technologią rozwijającą się dynamicznie. Coraz więcej producentów oferuje różne narzędzia, które umożliwiają ten rodzaj analiz. Przedsiębiorstwa zaczynają wykorzystywać technologię *data mining*, dostrzegając korzyści, jakie może przynieść im odkryta wiedza w prowadzonej działalności.

Eksploracja danych obejmuje różne techniki i algorytmy odkrywania wiedzy oraz predykcji, jak np. regresję, której celem jest odwzorowanie danych w wartości zmiennych predykcyjnych, które są liczbami rzeczywistymi. Odkrywanie charakterystyk jest problemem, którego rozwiązaniem są zwięzłe charakterystyki analizowanego zbioru danych. Na przykład moda na określony styl ubierania się może być zidentyfikowana przez zbiór reguł charakteryzujących. Dyskryminacja polega na odkrywaniu cech, które odróżniają wskazaną klasę obiektów od innych klas, np. zbiór reguł dyskryminujących może opisywać te cechy grupy klientów, które odróżniają daną grupę od innych.

Odrębną, aktualnie rozwijaną klasą algorytmów są narzędzia analizy i wizualizacji wzorców semantycznych w danych tekstowych. Umożliwiają one automatycznie wyodrębniać oraz wizualizować pewne stabilne wzorce i grupy wyrażeń, które często występują łącznie. Metody te pozwalają np. analizować komunikaty od klientów otrzymywane za pośrednictwem poczty elektronicznej lub centrów obsługi pod kątem poznania zależności między raportowanymi problemami a stopniem niezadowolenia klientów. W podobny sposób działają algorytmy wykrywające powiązania wartości dyskretnych oraz binarnych. Ich

zadaniem jest odkrywanie oraz wizualizacja wzorców i korelacji pomiędzy wartościami zmiennych.

Odkrywanie wzorców sekwencji polega na ustaleniu produktów z koszyka klienta z uwzględnieniem jego kolejnych zakupów. W bazie danych sklepu AGD możemy znaleźć produkty, które najczęściej poprzedzają kupno pralki. W sklepach RTV może być np. „sekwencja" telewizor – wideo. Z wzorca tego wynika, że klienci najpierw kupują telewizor, a po pewnym czasie wideo. Statystyczne wskaźniki wiarygodności oraz informacja o liczbie sprzedanych telewizorów pozwalając przewidzieć, ile wideo można sprzedać w najbliższym czasie. Możemy również zainteresować się klientami, którzy naruszają ten wzorzec i zastanowić się nad przyczynami.

Klasyfikacja danych jest procesem, którego celem jest znalezienie wspólnych cech charakterystycznych wśród obiektów bazy danych i przyporządkowanie ich do odpowiednich klas (grup), które pozwolą odróżnić je od pozostałych klas obiektów. Klasyfikacja jest znana i wykorzystywana w sztucznej inteligencji i uczeniu maszynowym. Wśród metod klasyfikacji możemy wyróżnić drzewa decyzyjne, których tworzenie odbywa się przez rekurencyjny podział zbioru na podzbiory – aż do uzyskania ich jednorodności ze względu na przynależność obiektów do klas, sieci neuronowe, algorytmy statystyczne. Wykorzystuje się ją do weryfikacji kredytobiorców, podziału pacjentów.

Technologia eksploracji danych może wspomagać procesy analizy problemów i podejmowania decyzji w rozmaitych sferach biznesu. Eksploracja danych ułatwia zarządzanie ryzykiem, wykrywanie oszustw i wyłudzeń, kontrolę jakości, ocenę konkurentów i ich strategii, inteligentne wyszukiwanie informacji, analizę danych tekstowych oraz internetowych zasobów informacyjnych. Narzędzia eksploracji danych umożliwiają także analizę i predykcję szeregów czasowych, analizę przepływów pieniężnych i tendencji rynkowych. Jednak do najlepiej znanych i głośnych zastosowań eksploracji danych należą zastosowania marketingowe.

Data mining, czyli drążenie danych, służy do wykrywana wzorców i powiązań pomiędzy danymi zawartymi w hurtowni danych. M.J.A. Berty i G. Linoff w książce *Data Mining Techniques for Marketing, Sales and Customer Support* podają następującą definicję: „*Data minig* jest to proces odkrywania i analizy, automatycznie lub półautomatycznie, dużych ilości danych w celu odkrywania znaczących wzorców i reguł" [Berry, Linoff 1997, s. 3]. *Data mining* jest wykorzystywane przede wszystkim do: klasyfikacji, estymacji, prognozowania, odkrywania reguł asocjacyjnych, grupowania na podstawie podobieństwa, analizy skupień, opisywania i wizualizacji danych.

Takich zależności może być „nieskończenie wiele"; dla użytkownika interesujące będą tylko niektóre z nich, i to w różnym stopniu. Nie można też idealnie zdefiniować, co kryje się pod określeniem „interesujące" [Nycz 2003, s. 351-361]. Dobre regularności mogą być uważane za: występujące również w nowych danych, spełniające narzucone warunki i preferencje użytkownika, wcześniej nieznane, acz

intuicyjne dla ekspertów, zrozumiałe i przejrzyste oraz przekładalne na wiedzę praktyczną.

Data mining stanowi zatem ogół technik znajdowania zależności pomiędzy danymi przez system. O ile analiza tradycyjna dzięki wiedzy i doświadczeniu użytkownika umożliwia mu tworzenie zestawień i porównań w samodzielny sposób, o tyle w *data mining* sam system potrafi odkrywać niektóre z takich zależności.

W miarę wzrostu ilości danych znacznie wzrasta złożoność obliczeń niezbędnych do wykonania analizy. Aby ograniczyć tę niedogodność, dokonuje się obliczeń, korzystając jedynie z reprezentatywnej próbki danych pochodzących ze zbioru źródłowego.

Przy drążeniu danych stosuje się następujące typy analiz: analizę szeregów czasowych – wykrywanie tendencji, estymację – szacowanie na podstawie analiz trendów, klasyfikację – przypisywanie przypadków do odpowiednich grup, łączenie – znajdywanie powiązań bez zależności formalnych, segmentację – wyodrębnianie podgrup o zbliżonych parametrach.

W wyniku przeprowadzonych procesów odkrywania wiedzy z hurtowni danych można zatem pozyskać wiedzę wcześniej nieznaną, która może okazać się bardzo użyteczna w procesie podejmowania decyzji.

5. Podsumowanie

Hurtownie danych są eksploatowane stosunkowo od niedawna. Operują na ogromnych zbiorach, głównie różnych bazach danych gromadzonych latami w przedsiębiorstwie. Może się tam znajdować się nieznana wcześniej albo nie uświadamiana wiedza. Narzędzia hurtowni mogą służyć do odkrycia tej wiedzy. Na szczególną uwagę zasługują zarówno narzędzia do przeprowadzania złożonych, wielowymiarowych analiz danych, jak też narzędzia *data mining* dostępne w hurtowni. Spożytkowanie odkrytej wiedzy zależy już jednak od człowieka, który podejmuje decyzje i za nie odpowiada.

Literatura

Advances in Knowledge Discovery and Data Mining, red. U. Fayyad, G. Piatetsky-Shapiro, MIT Press, Cambridge 1996.

Berry M., Linoff G., *Data Mining Techniques for Marketing, Sales and Customer Support*, J. Wiley, New York 1997.

Berry M., Linoff G., *Mastering Data Mining: The Art and Science of Customer Relationship Managemet*, J. Wiley, New York 2000.

Bębel B., Morzy T., *Projektowanie schematów logicznych dla magazynów danych*. Materiały z VI Konferencji PLOUG, Zakopane 2000.

Byrski M., *Data Mining w bazie Oracle 9i*. Materiały z VIII Konferencji PLOUG. Kościelisko 2002.

Chen M.S. Han J., Yu P.S., *Data Mining: An Overview from a Database Perspective*, IEEE Transactions on Knowledge and Data Engineering, 8(6) 1996: s. 866-883.

Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.
Han J. *Data Mining*, Kluwer Academic Publishers 1999
Inmon W.H., J.D. Welch, Glassey K.L., *Managing the Data Warehouse*, Wiley Comp. Publishing 1997.
Jakubczyc J., Nycz M., Smok B., *Hurtownie danych źródłem informacji decyzyjnych*, [w:] *Inteligentne systemy wspomagania decyzji w zarządzaniu. Nowa rola systemów informatycznych*, Prace Naukowe AE, Katowice 1998.
Kantardzic M., *Data Mining: Concepts, Models, Methods and Algorithms*, J.Wiley, New York 2002.
Morzy T., *Eksploracja danych: problemy i rozwiązania*. Materiały z V Konferencji PLOUG. Zakopane, październik 1999.
Morzy T., *Przetwarzanie danych w magazynach danych*, V Seminarium PLOUG, Warszawa 2002.
Nycz M. *Klasyfikacja danych w procesie inteligentnego pozyskiwania wiedzy z baz danych*, [w:] *Pozyskiwanie wiedzy i zarządzanie wiedzą*, red. M. Nycz, M. Owoc, Prace Naukowe Akademii Ekonomicznej nr 975, Wrocław 2003.
Nycz M., Smok B., *Ekstrakcja wiedzy z baz danych dla organizacji bazującej na wiedzy*, [w:] *Pozyskiwanie wiedzy i zarządzanie wiedzą*, red. M. Nycz, M. Owoc, Prace Naukowe Akademii Ekonomicznej nr 1011, Wrocław 2004.
Nycz M., Smok B., *Problemy związane z pozyskiwaniem wiedzy z baz danych*, [w:] *Pozyskiwanie wiedzy z baz danych*, Prace Naukowe Akademii Ekonomicznej nr 850, Wrocław 2000.
Nycz M., Smok B., *Wykorzystanie narzędzi Data Mining do odkrywania wiedzy wspomagającej decydenta*, [w:] *Komputerowo wspomagane zarządzanie*. Tom 2, red. R. Knosala, WNT, Warszawa 2004.

THE DATA ANALYSIS AND EXPLORATION METHODS

Summary

The paper has been devoted to the data warehouse exploration issues as a tool for realization multidimensional analysis as well as data mining. It consists of five parts. After short introduction, the idea of data warehouse has been described. The next part presents the warehouse methods that are especially useful in multidimensional analysis. Part four is devoted to data mining in warehouse. At the end short summary has been presented.

Dr inż. Małgorzata Nycz jest starszym wykładowcą w Katedrze Systemów Sztucznej Inteligencji Akademii Ekonomicznej we Wrocławiu
e-mail: malgorzata.nycz@ae.wroc.pl

Dr Barbara Smok jest adiunktem w Katedrze Systemów Sztucznej Inteligencji Akademii Ekonomicznej we Wrocławiu
e-mail: barbara.smok@ae.wroc.pl