Dariusz PAŁKA[1]
Piotr ZASKÓRSKI[2]

DATA MINING W PROCESACH DECYZYJNYCH

Streszczenie

W artykule dokonano opisu metod pozyskiwania wiedzy w modelach Data Mining stosowanych do wspomagania procesu podejmowania decyzji. Głównym założeniem jest próba wykorzystania do tego celu systemów klasy OLAP, jako systemów wielowymiarowych i wieloaspektowych drążeń informacji. Proces modelowania takich rozwiązań wymaga strukturalizacji i odniesienia do istniejącej bazy techniczno-technologicznej. Opracowanie prezentuje możliwości budowy modelu dla różnych klas organizacji oraz przedstawia możliwość adaptacji modelu data mining do analizy i zarządzania w procesie podejmowania decyzji. Przydatność modelu widziana może być szczególnie w aspekcie oceny możliwości wspomagania podejmowania decyzji związanych z planowaniem wykorzystania zasobów organizacji rozproszonych do przeciwdziałania skutkom zagrożeń. Nowoczesne koncepcje w zarządzaniu organizacją gospodarczą powinny eksponować platformę Internet, jako platformę ogólnie dostępną do komunikacji z otoczeniem.

Abstract

The present paper describes methods of knowledge absorption in the Data Mining models in order to support decision making processes. The main assumption is an effort to employ the OLAP systems as multidimensional and multiaspect data in drill down systems. The process of modelling such solutions requires structuring and referring to the existing technical and technological base. The paper presents possible options of model construction for different organisations and describes possible adaptation of the data mining model to the analysis and management of the decision making process. The applicability of the model may be viewed with respect to the analysis of the potential support of decision making with regard to the planning of utilisation of disperse organisations' resources in order to prevent the hazard effects. Modern concepts of economic organisation management should see the Internet as a widely accessible platform of communication with the environment.

1 WPROWADZENIE

Wspomaganie procesów podejmowania decyzji z wykorzystaniem technologii informacyjnych [11, 12] staje się w obecnej chwili wyzwaniem dla każdej organizacji. Szczególnie dotyczy to działania w warunkach niepewności i ryzyka. Takie sytuacje wiążą się

[1] Dr inż. Dariusz Pałka jest wykładowcą Warszawskiej Wyższej Szkoły Informatyki.
[2] Dr hab. inż. Piotr Zaskórski jest profesorem Warszawskiej Wyższej Szkoły Informatyki.

bezpośrednio z działaniem w stanach zagrożeń i kryzysów. Dostęp do wiarygodnej, jednolitej i wielowymiarowej informacji, możliwość koordynacji działań oraz usprawnienie procesów związanych z monitorowaniem i prognozowaniem niepożądanych skutków różnych zdarzeń i planowanie przeciwdziałania zagrożeniom – należą do istotniejszych obszarów decyzji. Skala biznesowych zastosowań różnorodnych technologii drążenia danych jest tak duża, że stanowią one *conditio sine qua non*[3] rozwoju i bezpieczeństwa współczesnego świata.

W celu opracowania odpowiedniego modelu systemu wspomagającego procesy decyzyjne niezbędna jest identyfikacja przedmiotu, procesu modelowania tej klasy systemów oraz identyfikacja danych a także ich źródeł pozyskiwania. Dokładna analiza zebranego materiału umożliwia wyodrębnienie jednolitych reguł w danych a tym samym wprowadza możliwość ich dalszego drążenia, z czym związane jest pojęcie Data Mining[4]. Zagłębianie danych [1] w data mining, jako proces analityczny, przeznaczony jest do badania dużych zasobów danych i zazwyczaj powiązanych z zagadnieniami gospodarczymi lub rynkowymi w poszukiwaniu regularnych wzorców oraz systematycznych współzależności pomiędzy zmiennymi, a następnie oceny wyników poprzez zastosowanie wykrytych wzorców do nowych podzbiorów danych. Finalnym celem data mining jest najczęściej przewidywanie zachowań różnorodnych zdarzeń i wspomaganie podejmowania decyzji.

2 BUDOWA MODELI DATA MINING

Data mining [7, 8] daje bezpośrednie korzyści w decyzjach biznesowych. Proces budowy modelu decyzyjnego w data mining składa się z czterech zasadniczych etapów[5]:

- wstępnej eksploracji,
- budowania modelu z określaniem wzorców,

[3]Tłumaczenie: warunek konieczny; http://en.wikipedia.org/wiki/Sine_qua_non, http://www.slownik-online.pl/index.php.

[4] Definicje „*Data Mining”*:

1. Nauka zajmująca się wydobywaniem informacji z dużych zbiorów danych lub baz danych (D. Hand, H. Mannila, P. Smyth: *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001).
2. Nietrywialne wydobywanie ukrytej, poprzednio nieznanej i potencjalnie użytecznej informacji z danych (W. Frawley, G. Piatetsky-Shapiro, C. Matheus: *Knowledge Discovery in Databases: An Overview*. AI Magazine, 1998).
3. Eksploracja danych jako jeden z etapów procesu odkrywania wiedzy z baz danych – www.wikipedia.pl.
4. Proces przekształcania danych w użyteczną wiedzę, począwszy od wprowadzenia danych lub pobrania ich z zewnętrznego źródła do utworzenia wynikowego raportu (Berry, M. J. A., Linoff G. S., *Mastering data mining*. New York: Wiley, 2000).

[5] Statis Soft: *Techniki zagłębiania danych, Internetowy podręcznik statystyki*. Kraków 2010.

- oceny i weryfikacji modelu,
- wdrożenia i stosowania modelu dla nowych danych, w celu uzyskania przewidywanych wartości lub klasyfikacji.

Eksploracja to pierwotny etap budowy modelu, który zaczyna się od przygotowania danych. Obejmuje przede wszystkim czyszczenie i przekształcanie, wydzielenie podzbiorów rekordów oraz wybór atrybutów danych, czego celem jest ograniczenie liczby analizowanych zmiennych do poziomu pozwalającego efektywnie wykonywać analizy (poziom ten zależy od stosowanych metod data mining). Po przygotowaniu danych dalszy przebieg eksploracji zależy od konkretnego problemu, który chcemy rozwiązać. Eksploracja może obejmować bardzo różne metody, od prostego wyboru predykatorów za pomocą regresji liniowej do wyrafinowanego badania danych różnymi metodami graficznymi i statystycznymi, którego celem jest wybranie najważniejszych cech i wyznaczenie ogólnej natury i stopnia złożoności modelu dla potrzeb drugiego etapu data mining.

Na etapie budowy i oceny modelu rozważane są różnorodne modele, po czym wybierany jest najlepszy z nich. Kryterium oceny jest jakość predykcji, czyli poprawność wyznaczania wartości modelowanej zmiennej i stabilność wyników dla różnych prób. Na pierwszy rzut oka wybór najlepszego modelu, może wydawać się dosyć prostym zadaniem, ale w praktyce czasami jest to skomplikowany proces. Istnieje wiele różnych metod oceny modeli i wyboru najlepszego z nich. Często stosuje się techniki bazujące na porównawczej ocenie modeli (ang. *competitive evaluation of models*) polegającej na stosowaniu poszczególnych metod dla tych samych zbiorów danych, a następnie wybraniu najlepszej z nich lub zbudowaniu modelu złożonego. Techniki oceny i łączenia modeli, które stanowią kluczową część w data mining sprowadzają się do następujących działań[6]:

- agregacji modeli (głosowanie i uśrednianie - ang. *bagging*),
- wzmacnianie modeli (losowanie adaptacyjne),
- łączeniem modeli (ang. *boosting*),
- uogólnianie modeli (ang. *stacking, stacked generalizations*),
- uczenie się modeli (ang. *meta-learning*).

Końcowym etapem budowania modelu data mining to wdrażanie i stosowanie, w którym stosuje się dla nowych danych model wytworzony i uznany za model najodpowiedniejszy. Gotowy model stosuje się w celu uzyskania przewidywanych wartości lub klasyfikacji [9].

[6] Tamże.

Dosyć istotnym pojęciem stosowanym w czasie budowy modelu data mining jest jego agregacja (ang. *bagging*), a także głosowanie (ang. *voting*) i uśrednianie (ang. *veraging*).

Agregacja modelu polega na zastosowaniu metody przewidywania wielu modeli tego samego typu uzyskanych dla różnych zbiorów uczących lub wielu modeli różnego typu uzyskanych dla tego samego zbioru danych. Modelowanie w tym przypadku zmiennej ciągłej tworzy procedurę zwaną uśrednianiem, a w przypadku zmiennych jakościowych dotyczących przedsięwzięć klasyfikacyjnych, stanowi głosowanie. Zastosowanie agregacji modeli umożliwia dokładniejsze i pewniejsze wyniki dla skomplikowanych zależności. Jest ono stosowane także, aby rozwiązać problem niestabilności i niewielkich rozbieżności wyników uzyskiwanych, gdy stosujemy skomplikowaną metodę dla małego zbioru danych. W przypadku uzyskiwania bardzo rozbieżnych wyników można zastosować metodę wzmacniania modeli.

Wzmacnianie modeli (ang. *boosting*) stosuje się w celu budowania kolejnych modeli dla danych i wyznaczenia wag dla modelu głównego. Pierwszy model budowany jest przy takich samych wagach wszystkich przypadków, a w kolejnych etapach wagi przypadków modyfikowane są tak, aby uzyskać dokładniejsze przewidywania dla tych przypadków, dla których wcześniejsze modele dawały błędne przewidywania. Wzmacnianie umożliwia także utworzenie sekwencji modeli, z których każdy jest wzorem w przewidywaniu dla przypadków, z którymi nie radziły sobie poprzedzające go modele.

Przygotowanie i weryfikacja danych jest nadzwyczaj ważna w procesie data mining. Analiza nadmiernych i zbędnych danych, bez rozwiązania powyższych problemów powoduje uzyskanie mylących wyników, szczególnie w czasie tworzenia modelu data mining stosowanym do przewidywania różnych działań i zdarzeń, czy też i przyszłości.

Redukcja danych w data mining dotyczy działań, których celem jest agregacja (amalgamacja) danych do postaci łatwiejszej do percepcji i przetwarzania [3]. Do redukcji danych wykorzystuje się proste techniki tabelaryczne, statystyki opisowe oraz bardziej wyrafinowane techniki np. analizę skupień i składowych głównych.

Wdrożenie i stosowanie w data mining oznacza zastosowanie wyników analizy dla nowych danych. Stosuje się go w tak zwanym predykcyjnym data mining i w klasyfikacji bezwzorcowej. Po uzyskaniu zadawalającego modelu lub podziału na segmenty, należy stosować te wyniki tak, aby szybko można było uzyskać przewidywane wartości lub przynależność do segmentu.

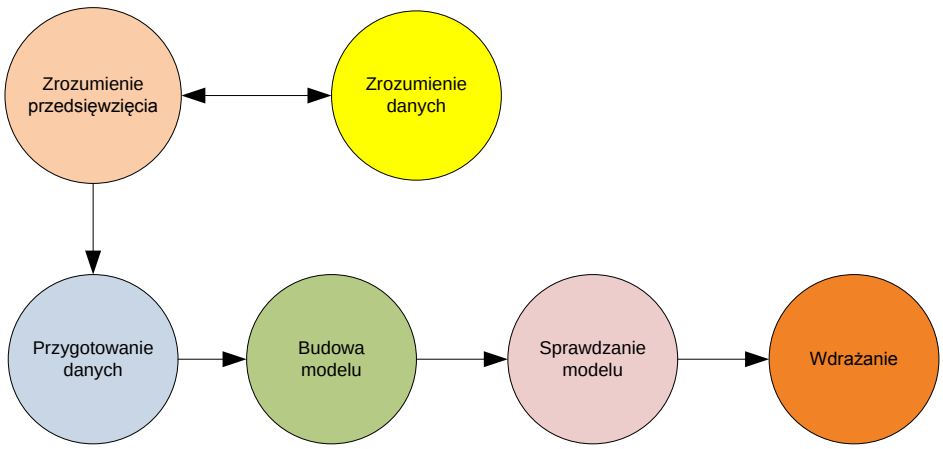
Bardzo popularną metodą budowy modelu jest technika drążenia danych (ang. *drill-down analysis*). W data mining polega ona na interakcyjnym badaniu danych, najczęściej dla dużych baz danych. Proces drążenia danych rozpoczyna wykonanie prostych przekrojów względem kilku zmiennych (czas, odległość, region). Dla każdej grupy wyznaczane są różnorodne statystyki, zestawienia oraz podsumowania. Na najniższym poziomie

określanym, jako dno, mamy dostęp do danych elementarnych, które to stanowią podstawowe źródło zasilania modelu.

W data mining często stosowany jest termin „uczącej się maszyny" rozumiany, jako ogólne określenie algorytmów dopasowywania modeli. W odróżnieniu od tradycyjnej statystycznej analizy danych, w której szacujemy parametry populacji metodami statystycznymi, w uczącej się maszynie data miningu nacisk nakładany jest na dokładność i użyteczność przewidywań lub na dokładny opis danych wynikowych.

W modelach predykcyjnych data mining stosuje się także metodę metauczenia w celu łączenia wyników wielu modeli w jeden model uogólniony. Technika ta jest w szczególności użyteczna, gdy modele są różnego typu [2]. Metodę tę często stosuje się w stosunku do wyników modelu zbiorczego, uzyskanego przez metauczenie, którą można stosować wielokrotnie. Jednak w praktyce zwiększa to ilość obliczeń, a uzyskiwana poprawa modeli jest coraz mniej znacząca.

Realizacja złożonych projektów data mining (zgłębiania danych) w organizacjach gospodarczych wymaga skoordynowanego wysiłku ekspertów, specjalistów i analityków różnych działów organizacji. W celu osiągnięcia oczekiwanego wyniku wymagane jest zastosowanie określonej metodyki, mogącej służyć, jako scenariusz, w jaki sposób należy zorganizować proces zbierania i analizy danych, rozpowszechniania wyników i sprawdzania korzyści z wdrażania projektu modelu.



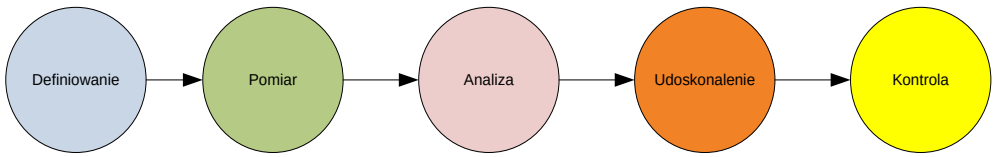
Rys. 1. Etapy budowy modeli data mining metodą CRISP

Jedną z głównych metod budowy modelu data mining jest metoda retrospektywnego drążenia danych.

Podstawową składową modeli retrospektywnych drążenia danych jest jednolita identyfikacja zasobów, procesów krytycznych i zagrożeń oraz ich źródeł w poszczególnych typach organizacji mających istotne znaczenie dla szeroko rozumianego bezpieczeństwa.

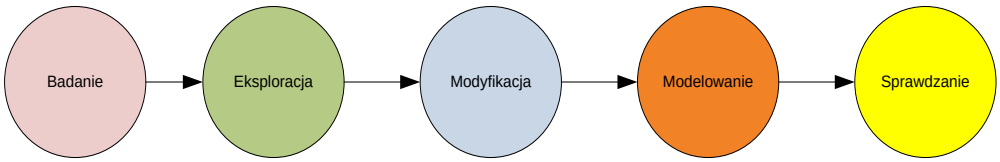
Inną z metod w budowy modelu data mining jest metoda CRISP (ang. *Cross-Industry Standard Process for Data Mining*). Metoda ta stała się powszechnie dostępnym standardem dla procesu data mining. Model ten postuluje ciąg sześciu etapów projektu data mining (rys.1).

Kolejną popularną metodą budowy modelu data mining jest metoda Sześć Sigma (ang. *Six Sigma*). Jest to metoda bazująca na strategii unikania wad w danych i problemów z ich jakością. Metoda ta proponuje pięć etapów budowy modelu (rys. 2).



Rys. 2. Etapy budowy modeli data mining metodą Six Sigma

Dosyć ciekawą metodą budowy modelu data mining jest metoda strategii SEMMA (*Sample, Explore, Modify, Model, Assess*) zaproponowana przez SAS Institute[7]. Metoda ta składa się z następujących etapów:



Rys. 3. Etapy budowy modeli data mining metodą SEMMA

Ogólnie należy stwierdzić, że wszystkie metody stosowane do budowy modeli data mining dotyczą sposobu przekształcania danych elementarnych na wiedzę oraz przedstawiają metodę jak udostępnić wiedzę, w takiej postaci, aby można było łatwo podejmować na jej podstawie decyzje strategiczne.

W podejściu data mining, drążąc zbierane dane, istotne jest znalezienie odpowiedzi na postawione problemy [4], ich rozwiązanie i przewidzenie hipotetycznego zdarzenia, ważnego z praktycznego punktu widzenia.

[7] SAS Institute to wiodący dostawca rozwiązań typu Business Intelligence oraz Data Mining.

Cały proces przekształcania danych w użyteczną wiedzę, począwszy od ich identyfikacji czy też pobrania z zewnętrznego źródła do utworzenia wynikowych decyzji wymaga zastosowania określonych metod.

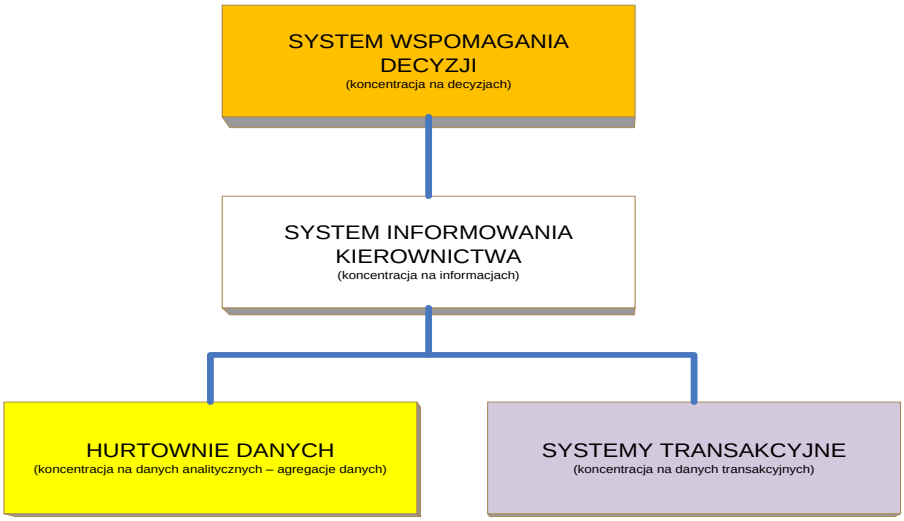
3 BUDOWA SYSTEMU WSPOMAGANIA PROCESU DECYZYJNEGO

Przyśpieszony rozwój elektroniki i różnorodność jej zastosowań stworzyły nowe możliwości przede wszystkim dotyczące automatyzacji procesów podejmowania decyzji poprzez zastosowanie rozbudowanych systemów informatycznych. Systemy te są nie tylko narzędziem gromadzenia i dostarczania danych podstawowych, lecz przede wszystkim stwarzają możliwość wspomagania procesów podejmowania precyzyjnych i kompletnych decyzji [5] wraz z wieloma agregacjami i prognozami.

Dynamika zmieniających się różnorodnych zdarzeń implikuje konieczność dokonywania szybkich analiz i podejmowania właściwych decyzji. Często analiza spływających informacji przekracza możliwości decyzyjne zespołów ludzkich. W tym miejscu w procesie decyzyjnym z pomocą przychodzą wyspecjalizowane Systemy Wspomagania Decyzji - SWD (ang. *Decision Support Systems*). Należy jednak zauważyć, że SWD nie zastępują człowieka, lecz pomagają rozwiązywać złożone problemy decyzyjne. Do głównych cech systemów tej klasy należą:

- komunikatywność, czyli przedstawienie informacji w sposób zrozumiały dla wykonawców,
- selektywność i interakcyjność informacji,
- integracja z danymi faktograficznymi,
- koncentracja na głównych decyzjach,
- szybkość reakcji na postawiony problem,
- łatwa i szybka manipulacja danymi.

SWD zwiększa wiedzę z zakresu zarządzania organizacją poprzez generowanie wielu możliwych wariantów decyzyjnych [14]. Aktualnie większość podejmowanych decyzji ma swoje podstawy analityczne tworzone przez wyspecjalizowane systemy informatyczne. Spływające informacje mają swoją wartość i wymagają odpowiedniego zarządzania podobnie jak inne zasoby każdej organizacji. W celu ograniczenia destrukcyjnej lawiny informacyjnej należy dążyć do selekcji informacji szczególnie o znaczeniu strategicznym dla określonych sytuacji [6] z uwzględnieniem czasu i miejsca ich pozyskiwania.



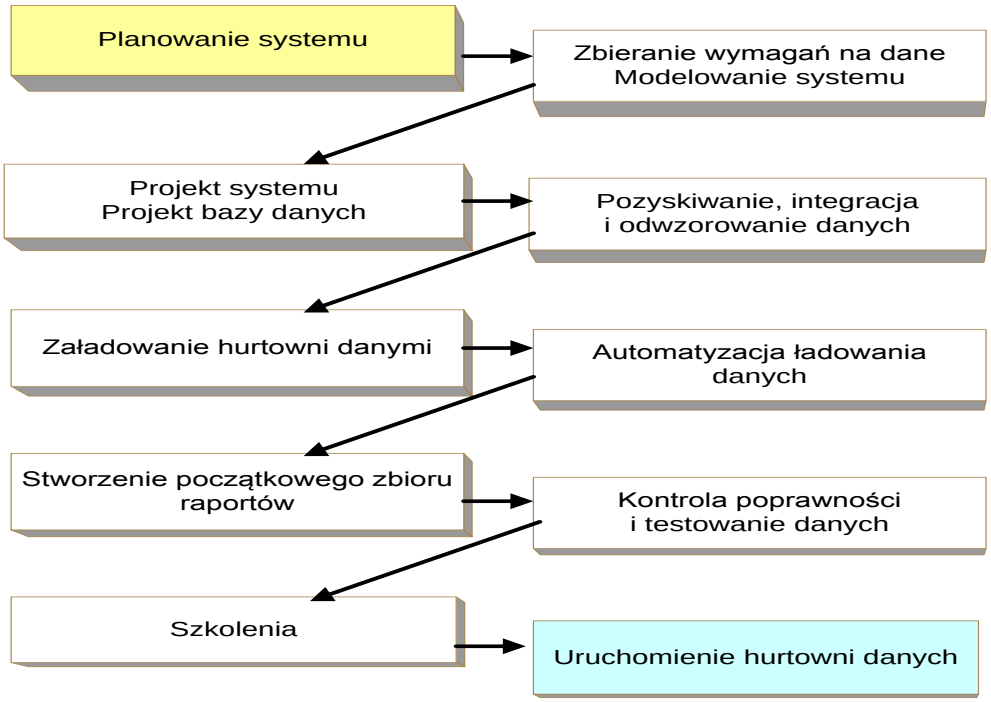
Rys. 4. Umiejscowienie Systemów Wspomagania Decyzji

Ważnym obszarem przygotowania decyzji jest odpowiednie czyszczenie i porządkowanie danych transakcyjnych (bieżących) oraz danych analitycznych (historycznych). Każda decyzja powinna mieć odniesienie do przeszłości szczególnie poprzez analizę stopnia analogii warunków i ograniczeń w konkretnym działaniu.

Dobrym modelem rozwiązań tej klasy problemów mogą być modele data mining budowane w oparciu o technologie OLAP (ang. *Online Analitycal Processing*) popularnie określanych mianem Hurtowni Danych (HD). HD stanowią system analityczny tworzony na podstawie danych pochodzących z heterogenicznych źródeł (różnorodnych systemów transakcyjnych OLTP - ang. *Online Transaction Processing*) i umożliwiający pełen zakres wielowymiarowego i wielokryterialnego drążenia danych tworząc tym samym określone modele data mining przydatne w procesie podejmowania decyzji.

HD mogą, więc wspomagać procesy analityczno-ocenowe i umożliwiać gromadzenie oraz przetwarzanie dużej ilości danych, jednoznacznie opisanych wymiarami wg potrzeb użytkownika. HD organizują i utrzymują dane analityczne, będące zwierciadłem zdarzeń w dłuższym przedziale czasowym. Układ tych danych jest zależny od potrzeb użytkownika, będącego decydentem określonego szczebla kierowania.

Istotnego znaczenia nabierają dziś modele data mining implementowane w środowisku sieciocentrycznym. Mogą mieć wówczas zastosowanie zarówno w systemach rozproszonych typu cloud computing jak i w wirtualnych architekturach hurtowni danych.



Rys. 5. Proces modelowania hurtowni danych jako bazy SWD [14]

Bieżący dostęp do zasobów danych analitycznych oraz danych transakcyjnych może być podstawą modelowania systemów wspomagania kierownictwa i dowodzenia a więc systemów wspomagania podejmowania decyzji (rys. 4). Tworzenie wektora różnorakich miar dla agregacji danych staje się źródłem dodatkowej wiedzy i decyzji planistycznych bazujących na różnorakich modelach prognostycznych.

Cały proces przygotowania i wdrażania narzędzi analityczno-decyzyjnych modelu data mining takich, jak hurtownie danych w organizacji gospodarczej jest złożony i czasochłonny. Wdrażając system wspomagający procesy podejmowania decyzji w dużym systemie działania o rozproszonej strukturze organizacyjnej z elementami rozmieszczonymi w różnych obszarach geograficznych - należy mieć na uwadze progresję poziomu trudności i czasu realizacji przedsięwzięcia. Dodatkowym utrudnieniem może być duża złożoność systemów transakcyjnych OLTP i wykorzystywanie przez daną organizację aplikacji dostarczanych przez różnych producentów oprogramowania czy systemów. Cały proces wdrożenia systemu wspomagającego podejmowanie decyzji charakteryzuje się pewnym modelowym cyklem życia. Cykl życia systemu SWD można odwzorować, jako sekwencję następujących po sobie kolejnych procesów (rys. 5).

W procesie planowania określa się plan całego przedsięwzięcia wraz z terminami realizacji poszczególnych elementów i wyznaczeniem czasu zakończenia realizacji roz-

wiązań [13]. Na tym etapie identyfikuje się również wymagane środki techniczne, a w szczególności wybór platformy sprzętowej i systemowej oraz rodzaj medium do transmisji danych i narzędzia ETL. Ustala się również zakres odpowiedzialności.

Zbieranie wymagań dotyczących danych oraz zakresu modelu jest sekwencją zdarzeń, w której identyfikuje się rzeczywiste potrzeby użytkowników wraz z obiektami informacyjnymi. Proces modelowania danych polega na zaprojektowaniu logicznego modelu bazy danych dla hurtowni danych (schemat bazy danych, klucze główne, atrybuty, rozmiar, tabele wymiarów oraz tabela faktów). Na tym etapie zaleca się wykorzystanie narzędzi wspomagających modelowanie i projektowanie systemów klasy CAISE (ang. *Computer Aided Information Systems Engineering*).

Modelowanie i projektowanie bazy danych związane jest z tworzeniem struktur danych do obsługi systemu DSS poprzez opracowanie modelu logicznego i powiązań obiektów bazy danych oraz algorytmów agregowania wg zadanych wymiarów i dla przyjętych miar agregacji. Jest to ważny etap modelowania systemu związany z identyfikacją obiektów oraz określeniem zbioru metadanych. Ten obszar modelowania związany jest przede wszystkim z nadawaniem nazw dla tabel i kolumn, przypisywaniem kluczy głównych oraz z tworzeniem procedur i funkcji składowanych, jeżeli jest to konieczne. Na tym etapie należy również zaplanować tryb dostępu do danych i kierować się wydajnością systemu poprzez opracowanie strategii indeksowania.

Pozyskiwanie, integracja i odwzorowanie danych jest jednym z najbardziej czasochłonnych etapów [6, 11]. Podczas tego etapu identyfikuje się potrzebne systemy źródłowe i dokonuje się wyboru najlepszych źródeł danych oraz przeprowadza się analizę danych w aspekcie ich integracji. Należy tu również zauważyć potrzebę ustalenia strategii integracji danych oraz przygotowania specyfikacji konwersji oraz opracowania modelu transformacyjnego danych źródłowych na dane docelowe, które mogą być przydatne do generowania decyzji w systemach działania antykryzysowego.

Załadowanie hurtowni danymi jest związane z modelowaniem procesu tworzenia procedur zapełniania bazy danych odpowiednimi danymi oraz oczyszczenia i uzupełnienia danych w lokalnych lub globalnych hurtowniach danych. Należy, więc stworzyć strategię umieszczania danych w hurtowni danych wraz z koncepcją procedur ładowania danych do hurtowni danych i modelem testowania narzędzi ETL.

Model automatyzacji ładowania danych wiąże się z automatyzacją procesów ETL. Podczas tego etapu należy zaplanować proces pobierania danych ze źródeł i zakres automatyzacji procesu transformacji danych. Ważnym elementem tego procesu jest plan automatyzacji procesu ładowania danych i testowania zautomatyzowanych procesów.

Generowanie początkowego zbioru raportów analityczno-planistycznych można już przeprowadzać w chwili umieszczenia pierwszych danych rzeczywistych w hurtowni da-

nych. Raporty można tworzyć również na podstawie wprowadzonych danych testowych. W celu wykonania procesu raportowania wykorzystać można narzędzia dostępu do danych. Na tym etapie należy poprawnie skonfigurować aplikacje dostępu do danych oraz dokonać weryfikacji zawartości raportów. Jeżeli testowe raporty odbiegają od przyjętych wymagań należy dokonać zmian w modelu transformacyjnym.

Kontrola poprawności i testowanie danych wiąże się z testowaniem już na etapie pobierania, przekształcania oraz ładowania, ale należy to również realizować kompleksowo. Dane na tym etapie mogą być sprawdzane pod względem formalnym i logicznym.

Złożoność procesów DSS tej klasy systemów wspomagania decyzji wymaga przeprowadzenia wielu szkoleń w celu uczynienia ich zrozumiałymi. Szkolenia powinny być procesem ciągłym według ustalonych programów szkoleniowych dla zarządzających i wykonawców, którzy będą korzystać z tej klasy systemów DSS. Użytkownicy systemu powinni być przeszkoleni podczas tego etapu cyklu życia DSS w zakresie narzędzi dostępu do danych, wykorzystania metadanych oraz samej koncepcji i idei modeli data mining.

Uruchomienie hurtowni danych jest etapem przygotowania stanowisk pracy dla użytkowników systemu oraz stworzenia procedur ich wspomagania. Ważnym procesem jest tutaj również opracowanie procedur wykrywania i usuwania problemów, a także opracowanie procedur rozbudowy aplikacji o nowe raporty i opracowanie procedur zarządzania metadanymi technicznymi i biznesowo-administracyjnymi.

4 MODEL SYSTEMU DECYZYJNEGO

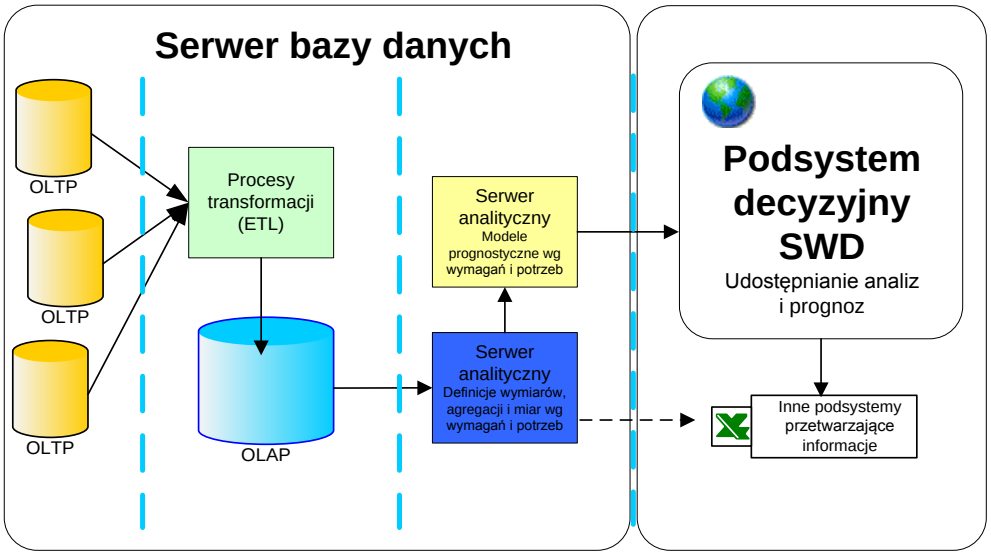
Idea modelu data mining wspomagającego proces decyzyjny bazujący na rozwiązaniach typu OLAP wiąże się z koniecznością określenia zakresu informacyjnego wraz z identyfikacją źródeł informacji w postaci funkcjonujących systemów klasy OLTP (rys.6). Systemy transakcyjne OLTP nie wnoszą istotnych ograniczeń technologicznych. Ważne jest, aby odpowiadały założonemu zakresowi informacyjnemu. Procesy transformacji danych z systemu OLTP do systemu analitycznego OLAP umożliwiają odpowiednią selekcję i przekształcenie danych do jednorodnej postaci zdeterminowanej wymiarami.

Komponent generowania analiz i dokonywania agregacji, a w tym definiowania miar specjalizowanych na bazie modeli prognostycznych wedle wymagań i potrzeb potencjalnych użytkowników decyduje o przydatności całego modelu. W komponencie tym następuje możliwość tworzenia uogólnień i generowania wiedzy poprzez planowanie przykładowo zasobów niezbędnych do przeciwdziałania określonym zagrożeniom. System może, więc udostępniać analizy i prognozy wspierające SWD.

Modelowanie i projektowanie [15] takich elementów SWD wymusza przeprowadzenie szeregu czynności, które mogą dać odpowiedź na pytania związane z zakresem informacyjnym przyszłych użytkowników hurtowni danych oraz ich oczekiwaniami wobec budowanego systemu. Zakłada się przy tym, że konieczne jest przeprowadzenie odpowiednich analiz istniejących systemów transakcyjnych pod kątem dostępności odpowiednich danych. Po przeprowadzeniu wieloaspektowych analiz i uzyskaniu niezbędnych informacji - podmiot modelujący określa model logiczny i model fizyczny struktur danych oraz model transformacyjny danych z systemów transakcyjnych wewnętrznych i ewentualnie zewnętrznych do modelu hurtowni danych.

Modelując hurtownię danych podmiot modelujący powinien uwzględniać ograniczenia, jakie występują i jakie mogą wystąpić w procesie wdrażania i użytkowania systemu.

W rozwiązaniach modelowych określa się również narzędzia, jakimi może dysponować podmiot modelujący, a w szczególności narzędzia wspomagające projektowanie systemów informatycznych typu CAISE oraz CASE (ang. *Computer Aided Software Engineering*). Istotną grupę modelowych narzędzi stanowi ETL (ang. *Extract, Transform, Load*), które wspomagają proces wydobywania danych dla potrzeb hurtowni danych. Głównym obszarem zastosowania narzędzi ETL jest pozyskiwanie danych ze źródeł zewnętrznych, przekształcanie danych i ich ładowanie do hurtowni danych. Baza technologiczna tej klasy rozwiązań udostępnia także systemy zarządzania bazami danych oraz narzędzia analityczne (ang. *Analysis Services, Business Intelligence Objects*).



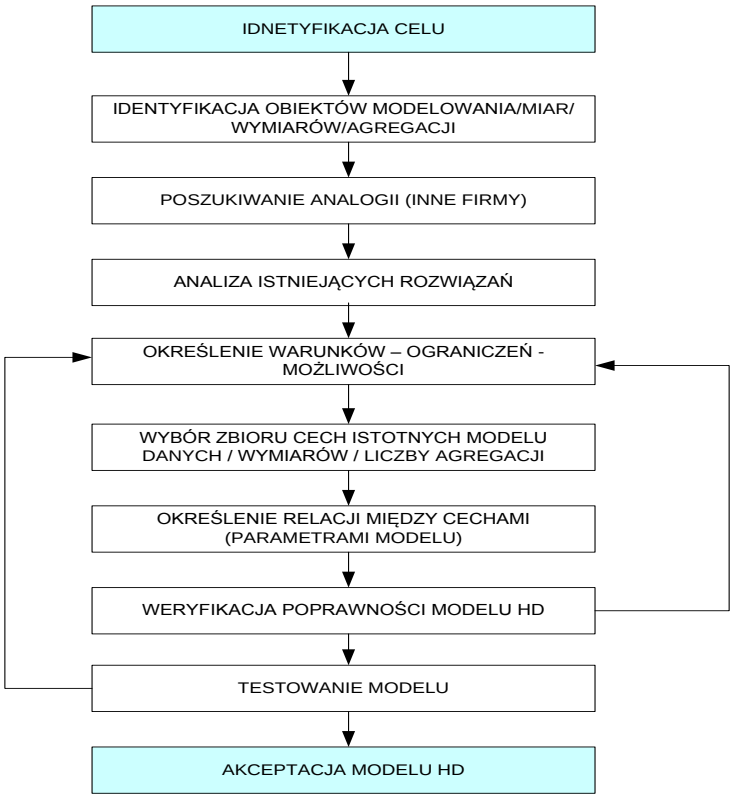
Rys. 6. Ogólny model hurtowni danych

Do głównych ograniczeń, jakie może napotkać podmiot modelujący należą ograniczenia finansowe, ograniczenia przepływności sieci oraz niedostępność wymaganych danych i ich wiarygodność. Ograniczeniem może być kolizja związana z wydajnością i wykorzystaniem pamięci.

Ograniczenia te powinny być uwzględniane w każdej fazie budowy modelu. Metodyka tworzenia hurtowni danych do wspomagania procesu decyzyjnego wskazuje na potrzebę dobrego odwzorowania potrzeb decydentów już na etapie opracowywania koncepcji systemu i jego modelowania logicznego. W fazie implementacji systemu i modelowania fizycznego mogą dominować ograniczenia techniczno-technologiczne. Faktyczną weryfikacją modelu jest testowanie integracyjne systemu. Ten etap warunkuje bezpośrednio oddanie do użytku i wdrożenie systemu.

Model logiczny systemu klasy SWD tworzony na bazie systemów klasy OLAP powinien określać funkcjonalność rozwiązań i identyfikować niezbędne dane, przykładowo o możliwościach i potrzebach związanych z przeciwdziałaniem w sytuacjach kryzysowych. Głównym jednak elementem tego komponentu jest ustalenie poziomu agregacji wprowadzanych danych i harmonogramu wprowadzania danych oraz wybór faktów i wymiarów. Schemat logiczny bazy danych (wg modelu gwiazdy, płatka śniegu) jest podstawą dalszego precyzowania modelu systemu SWD. Dla potrzeb rozwoju modelu ważnym jego elementem formalnym jest dokumentacja modelu logicznego.

Model fizyczny systemu zawierać powinien identyfikację różnych wariantów architektury programowo-sprzętowej oraz zasady instalacji i konfiguracji tej architektury, a także implementacji bazy danych. Dotyczy to przede wszystkim przeprowadzenia procesów ETL, konfiguracji serwera OLAP i procesów archiwizacji. Opracowanie modelu systemu typu SWD bazującego na rozwiązaniach HD wymaga skrupulatnego przestrzegania ustalonych faz jego tworzenia (rys. 7).



Rys. 7. Fazy tworzenia modelu hurtowni danych dla potrzeb SWD [12]

Tworzenie modelu hurtowni danych realizowane powinno być przede wszystkim poprzez identyfikację celu. Celem tego etapu jest określenie typów decyzji i ich zakresu informacyjnego w powiązaniu ze szczeblem zarządzania i klasą użytkownika. Identyfikacja obiektów modelowania oznacza określenie faktów, jakie będą identyfikowane, monitorowane i ewidencjonowane stosownie do określonych wymiarów oraz na wyznaczonym poziomie szczegółowości. Poszukiwanie analogii polega na zapoznaniu się, przez podmiot modelujący, jeżeli jest to możliwe, z podobnymi rozwiązaniami, funkcjonującymi w innych obszarach problemowych lub bazowanie na wzorcach z wcześniejszych doświadczeń. Analiza istniejących rozwiązań polega na przeprowadzeniu, przez podmiot modelujący, analizy funkcjonujących rozwiązań informatycznych oraz innych systemów dostępnych na rynku, przeznaczonych do budowy, jakich modeli. Celem etapu określenia warunków i ograniczeń jest identyfikacja występujących ograniczeń oraz tych, które mogą wystąpić w aspekcie wymagań stawianych modelowi. Wybór zbioru cech istotnych dla modelu danych hurtowni danych jest faza wyznaczenia atrybutów, jakimi charakteryzować się będą m.in. tabele wymiarów czy agregacje. Określenie relacji między cechami modelu polega na modelowaniu fizycznym hurtowni danych, natomiast

weryfikacją poprawności modelu jest przeprowadzenie weryfikacji powstałego modelu hurtowni danych w aspekcie postawionych wymagań. Jeżeli etap ten nie przebiegnie pomyślnie, podmiot modelujący powraca do etapu określenia ograniczeń i warunków. Celem etapu testowania modelu jest przeprowadzenie szeregu testów zależnych i niezależnych, np. przeprowadzenie próbnych analiz i raportów, testy wydajnościowe, testy bezpieczeństwa. Negatywna ocena testowa wymusza powrót do weryfikacji zbiorów istotnych cech modelu danych.

Wieloetapowa weryfikacja i doskonalenie modelu mogą być uproszczone przy zastosowaniu metodyki obiektowej. Odzwierciedlenie elementów i komponentów modelu z zachowaniem zasad generalizacji i polimorfizmu stwarza możliwość bezkolizyjnego wprowadzania udoskonaleń strukturalnych i funkcjonalnych. Model wspomagania procesów zarządzania i podejmowania decyzji powinien jednak mieć precyzyjne określone granice systemu i nie zawężać roli użytkownika oraz uwzględniać możliwość dynamicznego reagowania na zmiany potrzeb użytkownika. Kluczem do podnoszenia elastyczności i rozwoju modelu jest baza techniczno-technologiczna. Prezentowany model data mining uwzględnia różnorodność form i treści danych poczynając od dobrze uporządkowanych i jednoznacznych metadanych. Istotnym elementem modelu jest możliwość uzewnętrzniania wyników funkcji prognostyczno-planistycznych i analityczno-ocenowych również w środowisku zobrazowania graficznego.

Akceptacja modelu tej klasy systemów jest etapem kończącym proces modelowania hurtowni danych. Przyjmując jednak syndrom 98% w modelowaniu i projektowaniu systemów – można stwierdzić, że całość jest procesem otwartym, wymagającym stałego doskonalenia przedmiotowego modelu.

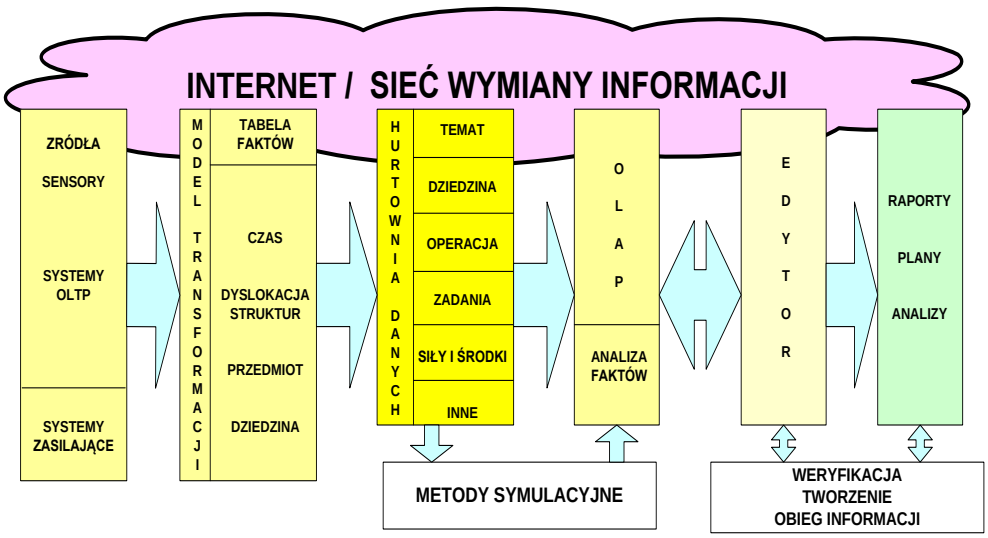
5 ZASTOSOWANIE DATA MINING W PROCESIE DECYZYJNYM

Budując model data mining dla wspierania procesu podejmowania decyzji trafną metodą jest wykorzystanie narzędzi hurtowni danych. Wieloprzekrojowa analiza danych modelu data mining [10] budowanych w środowisku OLAP może być skutecznym narzędziem długofalowej polityki w planowaniu strategicznym i operacyjnym.

Relacyjne bazy danych nie są wystarczającym rozwiązaniem dla systemów wspomagania decyzji. Specyfiką systemów analityczno-decyzyjnych jest to, że aby mogły sprawnie funkcjonować, potrzebują odpowiednio już przygotowanych (oczyszczonych, zagregowanych, przetransformowanych) danych.

Hurtownie danych organizują i utrzymują dane analityczne, będące zwierciadłem zdarzeń w dłuższym przedziale czasowym (rys. 8). Układ tych danych jest zależny od potrzeb użytkownika, będącego decydentem określonego szczebla kierowania. Szczegól-

nego znaczenia nabierają dziś data mining implementowane w środowisku sieciocentrycznym. Mogą mieć wówczas zastosowanie zarówno rozproszone jak i wirtualne modele architektury hurtowni danych. Instrument ten wspomaga procesy analityczno-oceno-we i umożliwia gromadzenie oraz przetwarzanie dużej ilości różnorodnych danych pochodzących z heterogenicznych źródeł, (z systemów klasy OLTP). Dane pochodzące z różnorodnych systemów informatycznych będą mogły być wieloaspektowo zwymiarowane wg potrzeb każdego typu użytkownika.



Rys. 8. Wydobywanie wiedzy w modelach Data Mining [11]

Narzędziem wspomagającym procesy decyzyjne a w szczególności planowanie – poprzez wykorzystanie wzorców użytkowych w obszarze planowania zasobów – mogą być systemy standardowe. Przykładem środowiska, w którym można implementować tej klasy modele są rozwiązania firmy SAS Institute oraz Oracle a także profesjonalne już dziś środowisko firmy Microsoft. Pakiet oprogramowania Microsoft SQL Server 2008 Analysis Services jest komponentem wielowymiarowego przetwarzania analitycznego w trybie on-line na serwerze Microsoft SQL Server 2008, integrującym różne struktury danych bezpośrednio powiązane z wymianą informacji poprzez Internet.

Raporty analityczne stanowią obraz wieloprzekrojowych analiz wspomagających procesy planistyczno-decyzyjne i pozwalają na szczegółowe zgłębianie informacji lub agregowanie wg parametrycznie określanych poziomów. Procesy planowania działań w każdej organizacji mogą być realizowane poprzez wykorzystanie odpowiednio zorganizowanych systemów informacyjnych. Wieloprzekrojowa analiza tych danych może być skutecznym narzędziem długofalowej polityki bezpieczeństwa w planowaniu strate-

gicznym i operacyjnym. Implementacja udostępniania danych na platformie internetowej stwarza możliwość integracji funkcjonalnej wielu organizacji uczestniczących w realizacji zadań w warunkach zagrożeń. Miary agregacji w takim modelu przybierać mogą wartości prognozy sytuacji oraz w konsekwencji, wartości planistyczne dla wcześniej zdefiniowanego zbioru zdarzeń i wynikających z tego klas i poziomów skutków.

Wymiana i eksploracja danych docelowo następować może w trybie transmisji danych z wykorzystaniem Internetu. Hurtownie danych porządkują proces informowania stosownie do potrzeb informacyjnych odzwierciedlanych w różnych przekrojach/wymiarach. Do podstawowych wymiarów zalicza się przede wszystkim czas, miejsce dyslokacji zdarzeń i zasobów, a także ich struktury przedmiotowe. Swoboda definiowania dowolnych wymiarów jest potwierdzeniem elastyczności modelu. Wymiary oraz zbiory faktów stanowią bazę do określania różnych poziomów agregacji. Agregacje złożone mogą odzwierciedlać bezpośrednio wartości prognostyczne. Źródłami danych mogą być dowolne organizacje danych utrzymywane w systemach bieżącego wspomagania, rejestrujących bieżące zdarzenia lub dane pochodzące z sensorów monitorujących bieżące możliwości działania, gdzie odzwierciedla się stan obiektu (wartość zasobów o określonej dyslokacji). W modelu data mining dokonuje się uporządkowania tych danych, przygotowując odpowiednie reprezentacje danych do bezpośredniej obsługi informacyjnej użytkownika w środowisku Internetu.

Ważnym aspektem zastosowań tej klasy modeli jest więc przeniesienie ich do środowiska sieci powszechnej. Bazowanie na Internecie daje możliwość obsługi informacyjnej wielu podmiotów organizujących i współuczestniczących w realizacji przedsięwzięć związanych z procesem reagowania kryzysowego. Szybka i bezpośrednia wymiana informacji oraz dostęp do zagregowanych zasobów informacyjnych mogą umożliwić koordynację działań oraz synchronizację procesów planistyczno-decyzyjnych. Dotyczy to przede wszystkim informacji o zasobach krytycznych związanych z sytuacjami kryzysowymi. Wieloaspektowe analizy i dostępne modele prognostyczne mogą być przedmiotem zainteresowania wielu organizacji na różnych szczeblach zarządzania.

6 PODSUMOWANIE

Wybór strategii monitorowania i zarządzania w procesach decyzyjnych oraz wdrożenie sprawnych struktur organizacyjnych wiąże się z dość znaczącymi nakładami. Wskazany model może służyć do wspomagania procesów analitycznych i planistyczno-decyzyjnych. Stąd też identyfikacja procesów decyzyjnych i poprawa trafności podejmowania decyzji wpływa na sprawne funkcjonowanie każdej organizacji poprzez kolejne wprowadzenie do modelu odpowiednich zmian i obserwowanie ich oddziaływania.

Zmiany mogą stać się siłą organizacji rozproszonych nadążających za dynamiką sytuacji kryzysowych.

W analizie porównawczej różnych strategii i metod działania należy kierować się również kryterium czasu i skróceniem cyklu decyzyjnego w sytuacjach krytycznych. Adaptacja niektórych koncepcji zarządzania w dłuższym horyzoncie czasowym może być ważnym punktem do analizy przyczynowej wprowadzania zmian strukturalnych i wprowadzenia nowych metod i technik zarządzania.

Analiza porównawcza metod i technik zarządzania według kryterium skuteczności wiąże się z identyfikacją ograniczeń i wymagań czasowych oraz kosztowych. Skuteczność modeli i metod zarządzania dużą ilością danych jest warunkowana często klasą stosowanych narzędzi. Aby organizacja mogła skutecznie realizować nałożone na nią zadania należy weryfikować i wdrażać nowoczesne koncepcje organizacji zasobów informacyjnych. Tak, więc wykorzystanie Internetu i modeli retrospektywnych data miningu może usprawnić istniejące proces podejmowania decyzji. Dobór metod i technik wspomagania zarządzania szczególnie w obszarze zapewnienia rozwoju działalności organizacji gospodarczej zależy od jej misji i wizji, a także od struktury wszystkich wewnętrznych procesów i zadań realizowanych przez organizację dla osiągnięcia zakładanych celów. Ważnym składnikiem jest potencjał organizacyjno-finansowy warunkujący zakres i możliwości współdziałania z otoczeniem bliższym i dalszym danej organizacji.

Rozwój technik i metod informacyjnych jest jednym z ważniejszych czynników doskonalenia zarządzania organizacjami rozproszonymi i daje możliwości wprowadzenia skutecznych usprawnień jakościowych w procesach decyzyjnych.

Literatura

1. Berry M. J. A., Linoff G. S.: *Mastering data mining*. New York, Wiley, 2000.
2. Frawley W., Piatetsky-Shapiro G., Matheus C.: *Knowledge Discovery in Databases: An Overview*. AI Magazine, 1998.
3. Han J., Kamber M., Pei J.: *Data Mining: Concepts and Techniques, Third Edition,,* The Morgan Kaufmann Series in Data Management Systems, San Francisco, CA, 2011.
4. Hastie T.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second edition, Stanford CA, 2009.
5. Pałka D., Zaskórski P., Żyto T.: *Risk in development and implementation of integrated systems*. VIII NATO Regional Conference on Military Communications and Information Systems. Gdynia, 2006.

6. Pałka D., Zaskórski P.: *e-Planowanie w logistyce organizacji gospodarczych*. VII Międzynarodowa Konferencja „Multimedia w Biznesie". Częstochowa, 2008.
7. Smyth P., Hand D., Mannila H., *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.
8. Statis Soft: *Techniki zagłębiania danych, Internetowy podręcznik statystyki*. Kraków 2010.
9. Sturm J.: *Hurtownie danych. SQL Server 7.0, Przewodnik techniczny*. MICROSOFT, 2000.
10. Surma J.: *Business Intelligence*. PWN, Warszawa, 2009.
11. Zaskórski P.: *Integracja procesów zarządzania organizacją*. Biuletyn WAT, 2006.
12. Zaskórski P.: *Strategie informacyjne w zarządzaniu organizacjami gospodarczymi*. WAT, Warszawa, 2005. s.275.
13. Zaskórski P., Suszek A.: *Zarządzanie procesami projektowo-wdrożeniowymi systemów bezpieczeństwa*. V Międzynarodowa Konferencja Bezpieczeństwa „Zarządzanie kryzysowe". Gdynia, 2007.
14. Zaskórski P., Zaskórski W.: *Retrospective models in operations planning.* VII NATO Regional Conference on Military Communications and Information Systems. Zegrze 2005.
15. Zaskórski P., Pałka D.: *Benchmarking w projektowaniu i wdrażaniu ZSyD*. XIV konferencja naukowa „Zautomatyzowane systemy dowodzenia i reagowania kryzysowego w procesie transformacji Sił Zbrojnych". Gdynia-Cetniewo, 2006.