

Jerzy Gołuchowski, Krzysztof Kania

Akademia Ekonomiczna w Katowicach

SYSTEM WSPOMAGANIA PROGNOZOWANIA Z OGRANICZENIAMI KALENDARZOWYMI

1. Wstęp

Zadania realizowane w działalności gospodarczej są zależne od czasu (zob. szerzej np. [KANI04], [KKG98]). Jednym z zadań bardzo często formułowanych w toku gospodarowania jest prognozowanie. Wraz ze zmianą liczby klientów lub zapotrzebowaniem na określone dobra i usługi zmienia się wielkość zasobów, które muszą mieć w dyspozycji ich dostawcy. Prognozowanie odgrywa szczególną rolę w przypadkach, gdy zasobów tych nie można zmagazynować (należą do nich np. energia i praca), ponieważ:

- niedobory zasobów mogą spowodować zakłócenia funkcjonowania organizacji lub poważne straty,
- nadmiar zasobów grozi marnotrawstwem (brak pracy dla pracowników, kary za niewykorzystane moce),
- szybka zmiana ilości posiadanych zasobów nie jest możliwa (lub jest kosztowna).

W takich przypadkach wymagane jest prognozowanie ze stosunkowo dużym wyprzedzeniem i dokładnością.

Niektóre z prognozowanych zjawisk lub zdarzeń wykazują specyficzną zmienność (okresowość), są związane z kalendarzem i przebiegają odmiennie w zależności od dnia tygodnia czy miesiąca. Można je określić jako uzależnione od kalendarza. Uzależnienie prognozowanych zjawisk, zdarzeń i procesów od kalendarza wynika najczęściej z uregulowań prawnych lub ogólnie przyjętych norm. Przykładem takiego zjawiska jest np. zróżnicowane zużycie mediów w gospodarstwach indywidualnych i przemyśle w niektórych dniach (np. przed różnymi świętami, w dni wolne od pracy). Podobny problem pojawia się także w planowaniu dostaw do dużych skle-

pów. Kolejnym przykładem jest wolumen i rodzaj dokumentów przyjmowanych w urzędach, bankach i pocztach w określone dni (płatności, wpłaty, wypłaty, przelewy) ze względu na terminy wypłat pensji (1. i 10. dzień każdego miesiąca). Ich specyfiką jest to, że na cykle (roczne, miesięczne, tygodniowe) zjawiska nakładają się układ dni świątecznych i dodatkowo wolnych od pracy. Prognozowanie w takich warunkach wymaga uwzględnienia tych czynników, które pomijają się w analizie zjawisk mających równomierny rozkład w czasie lub nieczułych na układ kalendarza.

W klasycznych metodach prognozowania ilościowego trudno jest uwzględnić układ kalendarza i zależność przebiegu procesów od rozkładu kalendarza. W takich przypadkach z pomocą przychodzi metoda jakościowa oparta na dużej ilości danych zgromadzonych w hurtowniach danych. Zadania tej klasy należą do zadań *business intelligence* (BI) i do ich wykonania wykorzystuje się różnorodne narzędzia statystyczne oraz techniki, takie jak sieci neuronowe czy analizę sekwencji (zob. np. [ADAM01], [HAKA01], [KOVI00]). W artykule zaprezentowano metodę wspomagania prognozowania opartego na danych zgromadzonych w hurtowni danych, uwzględniającą ograniczenia kalendarzowe. Metoda ta jest wykorzystywana w ING Banku Śląskim do prognozowania wolumenu przetwarzanych dokumentów bankowych oraz pracochłonności ich obróbki i jest wzywana w środowisku MS Excel z wykorzystaniem serwera MS SQL Server jako źródła danych.

2. Opis problemu

Oddziały ING BSK przyjmują dokumenty elektroniczne lub digitalizują (skanują) dokumenty tradycyjne (np. faktury, przelewy, opłaty ZUS, dokumenty podatkowe itp.). Zeskanowane dokumenty są weryfikowane przez system złożony z odpowiedniego oprogramowania i grupy osób o właściwych predyspozycjach (szybkość, dokładność, umiejętność skupienia itp.). Ilość pracy do wykonania jest zależna od wolumenu dokumentów oraz jego struktury (obróbka różnych typów dokumentów wymaga różnej ilości pracy). Ważnym ograniczeniem pracy zespołu jest konieczność wykonania pracy w ściśle określonym, nieprzekraczalnym terminie. Niezbędne jest zatem dokładne przewidywanie wolumenu dokumentów w poszczególnych dniach oraz prognozowanie pracochłonności ich obróbki i zaplanowanie niezbędnych zasobów pracy.

W ramach tworzonego w banku systemu klasy BI zbudowano moduł prognozowania ilości pracy i osób potrzebnych do przetwarzania dokumentów. Prognozowanie musiało uwzględniać następujące warunki:

- automatyczne pozyskiwanie danych o ilości i rodzaju przetworzonych dokumentów z plików transakcyjnych systemu przetwarzania dokumentów i możliwość ich grupowania według kryteriów użytkownika,
- stosunkowo niewielką ilość danych historycznych,
- duże (jednomiesięczne) wyprzedzenie umożliwiające zaplanowanie urlopów i zminimalizowanie liczby osób pozostających w gotowości do świadczenia pracy,

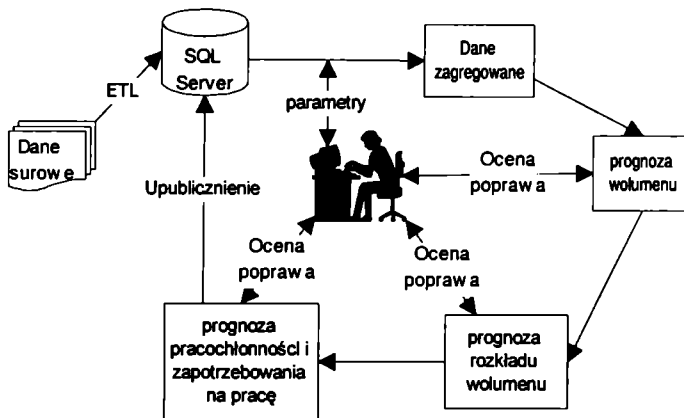
- duży wpływ czynników dodatkowych, takich jak np. zmiany opłat prowizyjnych, otwieranie/zamykanie oddziałów, zmiany rodzajów dokumentów, zmiany w ich strukturze, wprowadzanie nowych zasad elektronicznej wymiany informacji, powoduje, że konieczna jest współpraca systemu z doświadczonym użytkownikiem w celu weryfikacji uzyskanych prognoz.

3. Schemat procesu prognozowania pracochłonności obróbki dokumentów

Przetwarzanie odbywa się według następującego schematu (rys. 1):

- 1) pozyskanie danych z systemów transakcyjnych,
- 2) prognoza wolumenu dokumentów na cały prognozowany miesiąc,
- 3) prognoza rozkładu tego wolumenu na poszczególne dni robocze,
- 4) określenie ilości pracy i zamówienie określonej liczby pracowników,
- 5) bieżąca weryfikacja i ocena.

W prognozie wolumenu miesięcznego wykorzystywane są z dobrym skutkiem tradycyjne i często stosowane metody takie jak analiza trendu, analiza odchyłeń względnych i bezwzględnych oraz analiza podobieństwa przebiegów w analogicznych okresach poprzednich lat.



Rys. 1. Schemat przetwarzania w module prognozowania

Źródło: opracowanie własne.

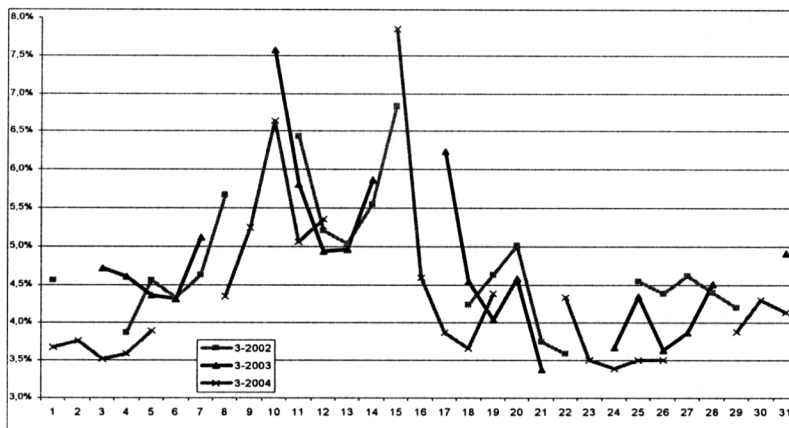
Najtrudniejszym problemem okazało się prognozowanie rozkładu wolumenu miesięcznego na poszczególne dni robocze. Rozłożenie wyznaczonego wolumenu na kolejne dni robocze miesiąca oparto na zastosowaniu:

- analizy rozkładu z przeszłości w miesiącach o podobnych charakterystykach,
- dodatkowych reguł korygujących uzyskany rozkład,
- możliwości swobodnego korygowania wyliczonego rozkładu przy założeniu przestrzegania nałożonych ograniczeń

W tym celu potrzebne było:

- znalezienie miesięcy podobnych do miesiąca prognozowanego,
- uwzględnienie występujących dni wolnych (ustawowo i dodatkowo),
- uzupełnienie metod statystycznych o zbiór reguł korygujących otrzymany rozkład przeciętny (średni).

Na rys. 2 przedstawiono rozkład ilości dokumentów w marcu w latach 2002-2004. Aby umożliwić porównywanie, dokonano normalizacji zjawiska. Wartości na osi pionowej wskazują, jaka część całego wolumenu miesięcznego przypadła na dany dzień. Widoczne jest, iż mimo porównywania analogicznych miesięcy kolejnych lat, rozkład zjawiska jest bardzo zróżnicowany. Źródłem tego zróżnicowania jest różny układ dni w tygodniu względem kolejnych dni w miesiącu, a co za tym idzie – różny układ najważniejszych dni względem dni w tygodniu. To zróżnicowanie powoduje, że niemożliwe jest posługiwanie się prostą analogią w rodzaju „rozkład w miesiącu z roku poprzedniego jest podobny do analogicznego miesiąca w roku bieżącym”. Powstaje zatem pytanie: na podstawie których miesięcy prowadzić prognozę rozkładu?



Rys. 2. Rozkład natężenia prac – kolejne dni wybranych miesięcy

Źródło: opracowanie własne.

4. Prognozowanie zjawisk zależnych od kalendarza

Aby odnaleźć miesiące podobne do miesiąca prognozowanego, wprowadzono pojęcie podobieństwa kalendarzowego miesiący. Jego obliczenie opiera się na położeniu dni kluczowych ze względu na rozkład prac względem dni roboczych i dni wolnych. W pracy banku w ciągu miesiąca (por. rys. 2) występują trzy takie dni tj. 10., 15. i 20. dzień każdego miesiąca. To właśnie położenie tych trzech dni względem dni wolnych i roboczych decyduje o obrazie rozkładu pracy również w kilku sąsiadujących dniach w miesiącu.

Do zapisania wiedzy o rozkładzie kalendarzowym zjawiska wykorzystano powiązane ze sobą macierze. Podstawowym powodem, dla którego zdecydowano się

na zapis informacji w postaci macierzy, a nie w postaci procedur lub funkcji, jest swobodny dostęp możliwość bezpośredniego operowania na wartościach i ingerencja użytkowników systemu w ich zawartość (np. wprowadzenie nowego dnia wolnego) bez użycia dodatkowego edytora lub specjalizowanego narzędzia.

Zawartość pierwszej macierzy związana jest z prostym spostrzeżeniem, że dzień tygodnia, od którego rozpoczyna się miesiąc, decyduje o rozkładzie miesiąca względem układu tygodnia aż do jego 28 dnia (zob. rys. 3). Aby określić, w jakim dniu tygodnia wypadnie dany dzień, wystarczy podać pierwszy dzień miesiąca. Miesiące mają jednakże różne długości. Dla analizowanego zjawiska nie jest to jednak przeszkodą, ponieważ żaden z dni krytycznych dla zjawiska nie przypadał na sam koniec miesiąca. Oznaczało to, że np. luty, mimo iż ma mniejszą liczbę dni, może być podobny do lipca.

	sty	lut	mar	kwi	maj	Cze	lip	sie	wrz	paź	lis	gru
2002	2	5	5	1	3	6	1	4	7	2	5	7
2003	3	6	6	2	4	7	2	5	1	3	6	1
2004	4	7	1	4	6	2	4	7	3	5	1	3
2005	6	2	2	5	7	3	5	1	4	6	2	4
2006	7	3	3	6	1	4	6	2	5	7	3	5
2007	1	4	4	7	2	5	7	3	6	1	4	6

Rys. 3. Fragment macierzy zawierającej numery dnia tygodnia, od którego rozpoczyna się dany miesiąc

Źródło: opracowanie własne.

Druga macierz (rys. 4) zawiera informację o podobieństwie między miesiącami w zależności od tego, w jakim dniu tygodnia rozpoczyna się dany miesiąc. W macierzy znajdują się wartości z przedziału $[0,1]$ wskazujące na stopień podobieństwa pomiędzy miesiącami. Na przykład element (1,2) macierzy oznacza, że rozkład zjawiska w miesiącach rozpoczynających się w poniedziałek jest podobny do rozkładu zjawiska w miesiącach rozpoczynających się we wtorek w stopniu 0,9. Element (4,7) o wartości 0,1 oznacza, że rozkłady dla miesięcy rozpoczynających się we czwartek i niedzielę są bardzo różne. Wartości dla tej macierzy pozyskano poprzez analizę kalendarza, wywiady z użytkownikami – ekspertami wspierane analizą graficzną danych historycznych oraz poprzez symulacje prognoz.

	1	2	3	4	5	6	7
1	1	0,9	0,8	0,3	0,3	0,1	0,3
2	0,9	1	0,9	0,6	0,6	0,1	0,1
3	0,8	0,9	1	0,6	0,3	0,1	0,1
4	0,3	0,6	0,6	1	0,9	0,1	0,1
5	0,3	0,6	0,3	0,9	1	0,3	0,1
6	0,1	0,1	0,1	0,1	0,3	1	0,9
7	0,3	0,1	0,1	0,1	0,1	0,9	1

Rys. 4. Macierz podobieństwa miesięcy

Źródło: opracowanie własne.

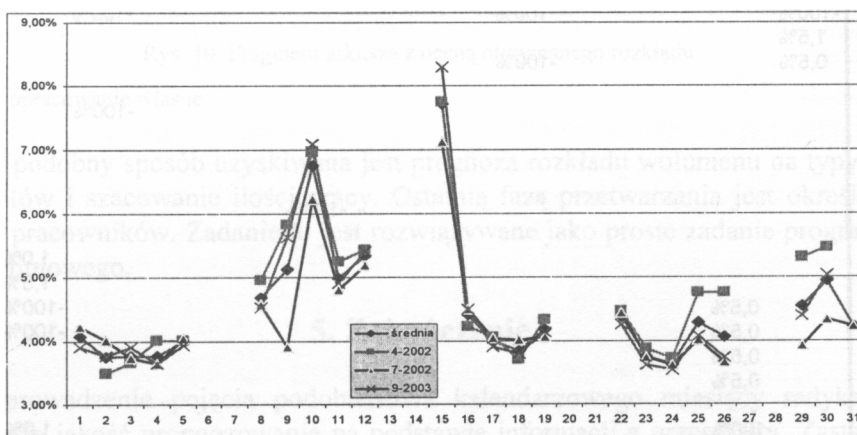
Zawartość obu macierzy służy do ustalenia zbioru miesięcy podobnych pod względem układu kalendarzowego do miesiąca prognozowanego. Na rys. 5 przedstawiono fragment arkusza z miesiącami podobnymi do prognozowanego marca 2004.

Do wykresu		Podaj odpowiedź		Próg 0,9
Suma: ile dok	mc-rok	7-2002	9-2003	rok 2002 mc styczeń - 0,9 rok 2002 mc kwiecień - 1 rok 2002 mc lipiec - 1 rok 2002 mc październik - 0,9 rok 2003 mc kwiecień - 0,9 rok 2003 mc lipiec - 0,9 rok 2003 mc wrzesień - 1 rok 2003 mc grudzień - 1
dzień	4-2002			
1	0,00%	4,23%	3,91%	
2	3,49%	4,01%	3,76%	
3	3,66%	3,73%	3,90%	
4	4,02%	3,64%	3,64%	
5	4,00%	4,04%	3,93%	
8	4,96%	4,57%	4,53%	
9	5,83%	3,90%	5,64%	
10	6,96%	6,22%	7,09%	
11	5,25%	4,81%	4,92%	
12	5,43%	5,19%	5,41%	
15	7,74%	7,12%	8,29%	
16	4,23%	4,49%	4,50%	
17	4,06%	4,04%	3,91%	

Rys. 5. Fragment arkusza z wybranymi miesiącami

Źródło: opracowanie własne.

O tym, które miesiące ostatecznie staną się podstawą prognozy, decyduje użytkownik. Z podanego zestawu do prognozy zostały wybrane przez użytkownika trzy miesiące ze wskaźnikiem podobieństwa 1; grudzień 2003, mimo, iż też miał wskaźnik podobieństwa równy 1, został odrzucony prawdopodobnie ze względu na dużą liczbę dni świątecznych. Na rys. 6 przedstawiono wykresy zjawiska dla wybranych miesięcy oraz wykres obliczonej średniej. Na wykresie widać, które okresy badanych miesięcy charakteryzują się stabilnością i prognozowanie jest wiarygodne (na wykresie dni od 16 do 24), które zaś charakteryzują się dużym rozrzu-



Rys. 6. Rozkład natężenia prac dla podobnych miesięcy (por. rys. 2)

Źródło: opracowanie własne.

tem (szczególnie koniec miesiąca). W dalszym ciągu procedury prognozy następuje weryfikacja otrzymanego rozkładu. Reguły weryfikujące rozkład przechowywane są w dwóch kolejnych macierzach.

Macierz trzecia i czwarta zawierają wartości dla reguł sugerujących, o ile należy skorygować wolumen przypadający na dany dzień wynikający z prognozy. Wszystkie stosowane reguły mają dokładnie dwa argumenty w przesłance i dokładnie jedną wartość w konkluzji. Umożliwiło to prostą implementację zapisu reguł w postaci dwuwymiarowej tablicy (rys. 7). Przykładowo element (9,Pt) trzeciej macierzy ma wartość 0,7%. Zapis ten odpowiada regule: Jeśli 9-ty dzień miesiąca wypada w piątek, to otrzymany rozkład należy podwyższyć o 0,7%. Z kolei wartości 0,5% dla ostatnich dni lutego z macierzy czwartej wynikają ze skumulowania się wolumenu rozłożonego w innych miesiącach na większą liczbę dni.

	Pn	Wt	Śr	Cz	Pt	So	N
1						-100%	-100%
...						-100%	-100%
9					0,7%	-100%	-100%
10	0,5%					-100%	-100%
11	0,7%					-100%	-100%
12						-100%	-100%
13						-100%	-100%
14					0,5%	-100%	-100%
15	1,0%					-100%	-100%
16	1,0%					-100%	-100%
...						-100%	-100%
31						-100%	-100%

Rys. 7. Fragment macierzy reguł korygujących rozkład w ujęciu tygodniowo-miesięcznym

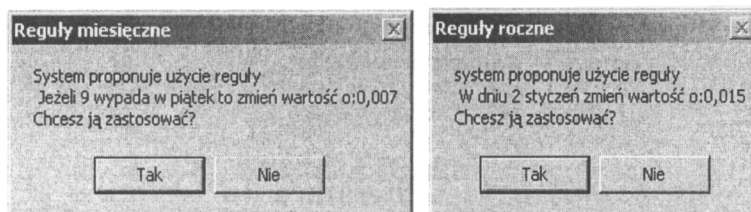
Źródło: opracowanie własne.

	Sty	Lut	Mar	Kwi	Maj	Cze	Lip	Się	Wrz	Paź	Lis	Gru
1	-100%				-100%						-100%	
2	1,5%											
3	0,5%				-100%							
...												
11											-100%	
12												
13												
14												
15								-100%				
...												
23												1,0%
24												1,5%
25		0,5%										-100%
26		0,5%										-100%
27		0,5%										
28		0,5%										
29												
30		-100%										1,0%
31		-100%		-100%		-100%			-100%		-100%	1,0%

Rys. 8. Fragment macierzy reguł korygujących rozkład w ujęciu miesięczno-rocznym

Źródło: opracowanie własne.

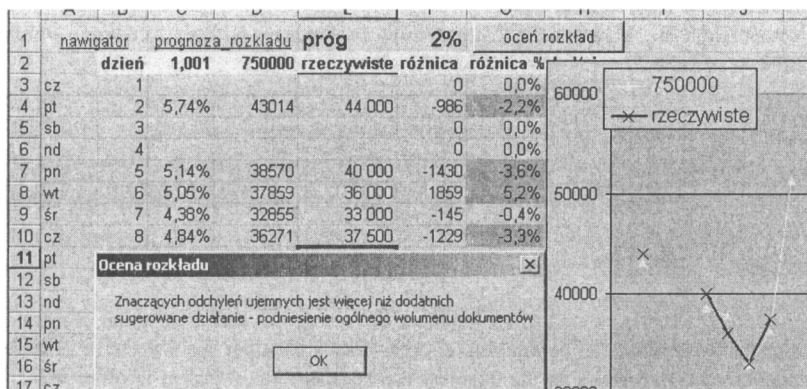
Informacja zawarta w macierzach jest przedstawiana użytkownikowi w postaci reguł (rys. 9).



Rys. 9. Przykłady reguł wygenerowanych przez system

Źródło: opracowanie własne.

W ostatniej fazie następuje ocena dopasowania rozkładu prognozowanego i rzeczywistego (rys. 10).



Rys. 10. Fragment arkusza z oceną otrzymanego rozkładu

Źródło: opracowanie własne.

W podobny sposób uzyskiwana jest prognoza rozkładu wolumenu na typy dokumentów i szacowanie ilości pracy. Ostatnią fazą przetwarzania jest określenie liczby pracowników. Zadanie to jest rozwiązywane jako proste zadanie programowania liniowego.

5. Zakończenie

Wprowadzenie pojęcia podobieństwa kalendarzowego miesiący radykalnie podniosło jakość prognozowania na podstawie informacji z przeszłości. Zastąpiło ono dotychczasowe wnioskowanie oparte na prostej analogii. Implementacja reguł o dokładnie dwóch elementach zawartych w przesłance reguły za pomocą arkusza kalkulacyjnego okazała się łatwa i efektywna.

Dodatkowo wprowadzono mechanizm przechowywania danych, parametrów oraz zbudowanych prognoz w bazie danych. Dało to możliwości przechowywania wielu prognoz i stosunkowo łatwego prowadzenia analiz w postaci *what-if*. Implementacja modułu spowodowała wzrost jakości, skrócenie czasu i zmniejszenie pracochłonności opracowywanych prognoz.

Literatura

- [ADAM01] Adamo J.M., *Data Mining for Association Rules and Sequential Patterns*, Springer-Verlag, New York 2001.
- [HAKA01] Han J., Kamber M., *Data Mining Concepts and Techniques*, Academic Press, 2001.
- [KKG98] Kania K., Kędzierski S., Gołuchowski J., *Zależności temporalne w modelowaniu i analizie procesów gospodarczych*, „Informatyka”, 3, 42, 1998.
- [KANI04] Kania K., *Temporalne bazy danych w systemach informatycznych zarządzania*, Wydawnictwo AE w Katowicach, 2004.
- [KOVI00] Kovalerchuk B., Vitayev E., *Data Mining in Finance*, Kluwer Academic Publishers, 2000.

FORECASTING SUPPORT SYSTEM WITH CALENDAR CONSTRAINTS

Summary

The paper addresses issues related to the problem of forecasting with a calendar constraints. In many cases calendar constraints can be omitted, in other ones a calendar has a decisive influence on the analysed phenomena. Especially in the business environment this problem is often seen. In the paper, by a particular example of business activity we present a new simply approach to resolve this problem.