

Big Data i Data Mining w polskim budownictwie

Dr inż. Marcin Gajzler, Instytut Budownictwa, Politechnika Poznańska

1. Wprowadzenie

Pojęcie Big Data odnosi się do bardzo dużych i zróżnicowanych zasobów danych, jakie są w dyspozycji. Termin ten w ostatnim dwudziestoleciu upowszechnił się, zaś Big Data stało się cenionym zasobem, który poddany analizie stanowi częstokroć podstawę podejmowania decyzji, a także zarządzania strategicznego. Analizując bogatą już literaturę [2, 10, 11, 13], ale także praktyczne aplikacje oraz szeroką ofertę szkoleń i seminariów zauważa się, że pojęcie Big Data występuje w procesach biznesowych, w zarządzaniu, w handlu, produkcji i usługach. Metodą pozyskiwania potencjalnie cennych informacji z dużych zasobów danych jest grupa technik określana jako Data Mining [1]. Pozwala ona w relatywnie krótkim czasie na przeszukanie dostępnego zasobu danych i zdobycia informacji, których pozyskanie poprzez tradycyjne analizy byłoby kosztowne i czasochłonne, a nawet niemożliwe. Celem artykułu jest udzielenie odpowiedzi, czy również sektor polskiego budownictwa wykazuje zainteresowanie analizami Big Data, a przede wszystkim czy w sektorze budownictwa występują zasoby danych, dla których termin Big Data jest właściwy.

2. Big Data

Wraz z rozwojem komputeryzacji, możliwościami trwałego gromadzenia i przechowywania danych, a także w związku z coraz bardziej rozwiniętym rynkiem usług, gdzie konkurencyjność organizacji stale wzrasta, nastąpił rozwój informatycznych systemów zarządzania. Od prostych systemów transakcyjnych (TPS Transaction Processing Systems) po inteligentne i zintegrowane systemy zarządzania organizacją (OSS Organization Support Systems) [14]. Na trendzie tej ewolucji, w odpowiednich warunkach wykształcił się kierunek określany jako Big Data, który odpowiada za gromadzenie i analizę dużych i zróżnicowanych zasobów danych, w celu dostarczenia przydatnych, w tym ukierunkowanych i nieukierunkowanych informacji. Co istotne, wspomniane zasoby danych ciągle się rozrastają, albowiem rozwój technologii mobilnych umożliwia przesyłanie i ciągłe gromadzenie danych w czasie rzeczywistym. Powoduje to, że typowe bazy danych, jakie współlistniały z systemami transakcyjnymi w znaczący sposób różnią się od współczesnych zasobów danych i zastosowanie technik analizy tych danych wymaga współczesnego i zautomatyzowanego wobec tak ogromnych rozmiarów danych podejścia. Jako przykład takiego podejścia odpowiadającego

wymaganiom dużych i zróżnicowanych zasobów danych należy wskazać zbiór technik Data Mining. Są to techniki określane jako eksploracyjne pozwalające na odkrywanie wiedzy i wywodzące się z połączenia technik baz danych, metod statystyki oraz metod sztucznej inteligencji. Dzięki zastosowaniu technik komputerowych współczesna analiza dużego i zróżnicowanego zasobu danych z wykorzystaniem technik Data Mining nie stanowi obecnie problemu czy wyzwania. Co ciekawe, przydatność dużych zasobów wiedzy poddanych analizie zasadniczo nie ulega istotnej redukcji. Wynika to z faktu, że w podejściu Data Mining możemy wyróżnić podejście ukierunkowane, które stawia sobie za cel pozyskanie określonej informacji i wiedzy, co tylko potencjalnie wyczerpuje przydatność zbioru po wykonanej analizie. Obok podejścia ukierunkowanego wyróżnia się nieukierunkowane, które opiera się o analizy nawet już wcześniej eksplorowanych danych w celu pozyskania nowych informacji czy poznania nowych prawidłowości, które zależnie od potrzeb organizacji mogą okazać się przydatne.

3. Dostępne zasoby danych

Rozwój technik komputerowych pozwolił na gromadzenie wielu danych będących rezultatem realizacji procesów, takich jak: transakcje, procesy produkcyjne, monitoring jakości, logistyka, kontrola stanu itp. Oprócz tego, co jest akcentowane w szczególności poza naukami technicznymi, zasoby sieci internetu stanowią ogromny zasób danych, który traktuje się jako swoistą kopalnię informacji i wiedzy o zjawiskach i o samych użytkownikach. Różnego rodzaju portale społecznościowe i dane tam zawarte podlegają wnikliwym analizom pozwalającym na odkrycie informacji, np. o preferencjach użytkowników. To pozwala dopasować prezentowane treści do odbiorcy, a nawet przygotować ofertę, którą być może użytkownik będzie zainteresowany, zanim jeszcze podejmie jakiegokolwiek własne działania.

W Big Data można wyróżnić podstawowy model opierający się na pięciu atrybutach [3, 12], takich jak:

- wielkość (Volume),
- prędkość (Velocity),
- różnorodność (Variety),
- wartość (Value),
- wiarygodność (Veracity).

Wielkość to atrybut dotyczący zasobności zbiorów danych. Jak wspomniano wcześniej, rozwój technik komputerowych w tym trwałych nośników, a także rozwiązań rozproszonych

(chmury) umożliwiają gromadzenie danych, nawet takich, co do których aktualnie nie ma uzasadnienia ich niezbędności. O ile kilkadziesiąt lat temu dane były gromadzone w formie formularzy papierowych i przechowywane w archiwach, o tyle aktualnie dominują nośniki elektroniczne, znacznie przekraczające wolumen niejednego archiwum.

Prędkość to atrybut związany obecnie, znów dzięki rozwojowi technologii, w szczególności technologii monitoringu oraz bezprzewodowego przesyłu danych, z dostępnością danych w czasie rzeczywistym. Są to dane o charakterze dynamicznym wynikające z monitoringu stanu obiektu czy zjawiska.

Różnorodność ma związek z formą, a także strukturą gromadzonych danych. Ze względu na różnorodność źródeł danych wymagają one często wstępnej transformacji i organizacji stanowiącej element zasadniczej analizy Big Data. Należy uwzględnić, że dane nie muszą mieć charakteru ilościowego, ale dość powszechne i cenne są dane o charakterze jakościowym. Istnieją odrębne techniki ich analizy, co pozwala wykorzystać je w zdobywaniu informacji i wiedzy.

Wartość to atrybut opisujący znaczenie danych dla analizy. Wydaje się być to atrybut subiektywny pomimo metod pozwalających na określanie siły wpływu zmiennych na zależną. Wobec prowadzenia tzw. analiz nieukierunkowanych praktycznie każde dane mają potencjałe znaczenie. Jednak gdy pozostaje się w analizach ukierunkowanych i wybiera określone narzędzia (np. sztuczne sieci neuronowe), wstępna selekcja danych wydaje się być kluczowa.

Wiarygodność wydaje się być jednym z ważniejszych atrybutów istotnych dla wyników analizy. O ile dane dotyczą procesów stale rejestrowanych (np. transakcje kasowe, pomiary stanu elementów obiektu, monitoring zjawiska), o tyle m.in. w budownictwie mogą wystąpić źródła danych i same dane, co do których wiarygodności istnieją uzasadnione obawy (np. ze względu na niejawność procesu lub czynniki pozaformalne). Analiza danych o niskiej wiarygodności powoduje tym samym niską wiarygodność jej wyników i ograniczoną ich przydatność.

Rozpatrując duże zasoby danych odpowiadające atrybutom Big Data z pewnością można zidentyfikować je w sektorze budownictwa. Zasoby te mają różny charakter. Począwszy od danych ilościowych o określonej nadanej strukturze, jakimi są przykładowo katalogi nakładów (np. KNR, KNNR), ilościowe zestawienia materiałowe dla obiektów, przez zasoby tekstowe opisujące zjawisko czy stan obiektu (specyfikacje techniczne, opinie i ekspertyzy dotyczące stanu technicznego, protokoły pokontrolne wykonania robót), dynamiczne dane pomiarowe (pozyskiwane w czasie rzeczywistym z wykorzystaniem technik bezprzewodowych pomiary stanu procesu, elementu – np. dane dotyczące zmian temperatury w betonie, czasy realizacji procesów technologicznych podczas betonowania, dane z monitoringu obciążenia elementu) – aż do wykorzystania modelowania BIM i zawartej tam struktury danych elementów modelu.

Oprócz tego w budownictwie występują typowe zasoby funkcjonujące w innych dziedzinach gospodarki – np. dane księgowe przedsiębiorstw, dane klientów i kontrahentów czy dane transakcyjne. Mają one również istotne znaczenie i pozwalają wspomagać procesy zarządzania zarówno strategicznego, jak i operacyjnego.

Niestety wciąż zauważa się, że świadomość użytkowników czy dysponentów tych zasobów, co do możliwości ich wykorzystania, nie jest w pełni rozwinięta. Wyniki badań w tym zakresie opierające się na ankiecie zostaną zaprezentowane w dalszej części.

4. Znajomość Big Data w sektorze polskiego budownictwa

Praktyczne wykorzystanie dużych zasobów danych uzależnione jest w pierwszej kolejności od znajomości zagadnienia oraz dostępności narzędzi analizy. W związku z tym pod koniec 2020 roku przygotowano ankietę, którą rozesłano drogą elektroniczną do 58 przedsiębiorstw sektora budowlanego działających na terenie kraju (w tym do różnych działów w przypadku dużych przedsiębiorstw). Ankieta zawierała 12 pytań dotyczących znajomości zagadnień Big Data, możliwości wykorzystania zasobów danych i stanu faktycznego w przedsiębiorstwie. Zwrot ankiet używano na poziomie ok. 77,6%. Jednym z pierwszych pytań zawartych w ankiecie było pytanie dotyczące znajomości zagadnień Big Data wśród osób średniego szczebla zarządczego. Aż 41 respondentów wskazało na znajomość Big Data, przy czym jedynie 7 potwierdziło znajomość zagadnień na podstawie dedykowanych szkoleń czy warsztatów. Można na podstawie tego sformułować niekategoryczny wniosek o braku potrzeby szczegółowego poznania Big Data i tym samym braku świadomości i perspektyw zastosowania analiz Big Data w praktyce. W opinii autora zagadnienia Big Data wymagają poznania szczegółowych metodyk, a sama ogólna znajomość pojęcia jest zdecydowanie niewystarczająca do praktycznego wykorzystania. W kolejnych punktach ankiety zapytano o dostępność zasobów Big Data w przedsiębiorstwie i na zewnątrz przedsiębiorstwa, wskazując przy tym na przykłady takich zasobów. Udzielane odpowiedzi wskazywały na dostępność dużych zasobów danych w przedsiębiorstwach, co winno dawać podstawy do ich analiz. Jednym z kolejnych punktów ankiety było pytanie o dostępność narzędzi analizy Big Data. W tym przypadku jedynie jedno przedsiębiorstwo udzieliło odpowiedzi pozytywnej. Zauważa się, że w kontekście nakładów na zakup oprogramowania istnieją alternatywne sposoby wykonania analizy i tego dotyczyło kolejne pytanie. Okazało się, że 3 przedsiębiorstwa zlecały analizy Big Data do podmiotów zewnętrznych. Z tego wynika, że jedynie ok. 9% przedsiębiorstw miało styczność z praktycznymi analizami Big Data. W oparciu o wykonane badania ankietowe stwierdza się, że były to przedsiębiorstwa

deweloperskie, które w ten sposób badały preferencje swoich potencjalnych klientów, chcąc zwiększyć swoją konkurencyjność i dostosować ofertę.

Przeprowadzone badanie ankietowe niestety wykazało, że w badanej próbie znikome jest praktyczne wykorzystanie Big Data pomimo dysponowania odpowiednimi zasobami danych. Dotyczy to zarówno właściwych dla sektora budownictwa zasobów, jak i typowych – np. księgowych.

Na podstawie tego można sformułować wniosek, że sektor polskiego budownictwa jest jeszcze w przededniu „rewolucji informacyjnej” opartej na Big Data. W opinii autora być może w niedalekiej przyszłości wzrośnie świadomość dysponowania zasobami danych, jakie nie zostały jeszcze w całości wyeksplorowane i cały czas dysponują potencjałem informacyjnym, który przyniesie wymierną korzyść dla organizacji.

5. Możliwości wykorzystania analiz Big Data

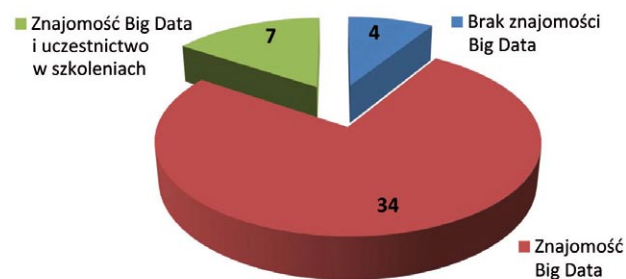
Jak wynika z przeprowadzonych badań ankietowych, upowszechnienie Big Data w sektorze polskiego budownictwa nie osiągnęło jeszcze wysokiego poziomu. Stan ten potwierdzają również badania opisane w [8, 9]. Tymczasem aplikacyjność do wielu zagadnień obecnych w sektorze budownictwa jest w oparciu o doświadczenia innych branż znana od lat.

Przeprowadzone badanie ankietowe nie pozwoliło precyzyjnie określić zakresu prowadzonych przez przedsiębiorstwa analiz Big Data. Z informacji uzyskanych można jedynie wnioskować, że prowadzone analizy miały na celu rozpoznanie preferencji zakupowych klientów w związku ze sprzedażą przez przedsiębiorstwa wykonanych lokali mieszkalnych. Autor zauważa, że w [6] zaprezentował możliwości analiz Data Mining o podobnym kierunku. W [6] zaproponowano analizę Data Mining dla zasobów danych pozyskanych z internetowych portali ogłoszeniowych poświęconych nieruchomościom, witryn internetowych firm deweloperskich, a także ofert biur obrotu nieruchomościami. Celem analizy było określenie istotnych atrybutów wpływających na zainteresowanie się ofertą przez potencjalnych klientów, w szczególności określenie atrybutów lokalizacji. Była to analiza ukierunkowana – zmierzająca do ściśle określonego celu. W celu rozwiązania problemu i wskazania istotnych przy wyborze lokalizacji atrybutów wykorzystano w analizie drzewa klasyfikacyjne C&RT (Classification and Regression Tree). Uzyskane wówczas wyniki pozwoliły zwrócić uwagę na takie atrybuty lokalizacji, jak odległość od obiektów infrastruktury społecznej (zdrowie, oświata i rekreacja), a także o nieznacznie niższej istotności – odległość od handlu i usług oraz odległość od sieci komunikacji.

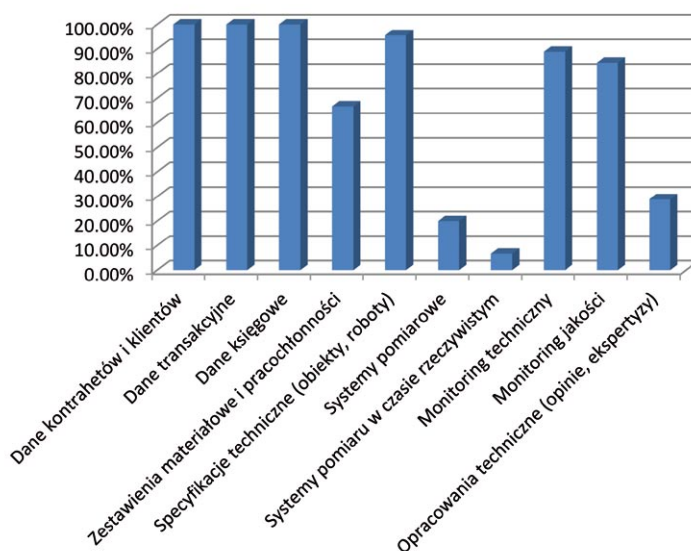
Analizę o podobnym zakresie zaprezentowano również w [5], przy czym źródło danych stanowiła pozyskana od firm

deweloperskich charakterystyka poziomu sprzedaży lokali mieszkalnych (częściowo dane transakcyjne). W analizie uwzględniono również czynniki lokalizacyjne, zaś samą analizę można również określić mianem ukierunkowanej. Wyniki badania wskazały na ciekawą, ale i wydaje się oczywistą okoliczność. Z wyników analizy można wywnioskować, że to wcale nie atrybut ceny zakupu lokalu jest najistotniejszy w decyzji potencjalnych klientów. Atrybutami o wyższej istotności była odległość od hałaśliwych ulic, odległość od lotniska, dostępność infrastruktury społecznej i edukacyjnej.

Inne aplikacje autora dotyczyły systemów wspomagania decyzji technologicznych [4, 5]. Problemem, jaki został rozwiązany poprzez analizę Big Data, było grupowanie dostępnych na rynku rozwiązań materiałowych z punktu widzenia zastosowań. Było to przydatne z punktu widzenia doboru odpowiedniego materiału. Cechą szczególną prowadzonych analiz była specyfika zasobu danych. Poprzez podejście Text Mining w analizie wykorzystano bowiem dane tekstowe – fragmenty kart technicznych różnych rozwiązań materiałowych. Zwrócono wówczas uwagę na pewne

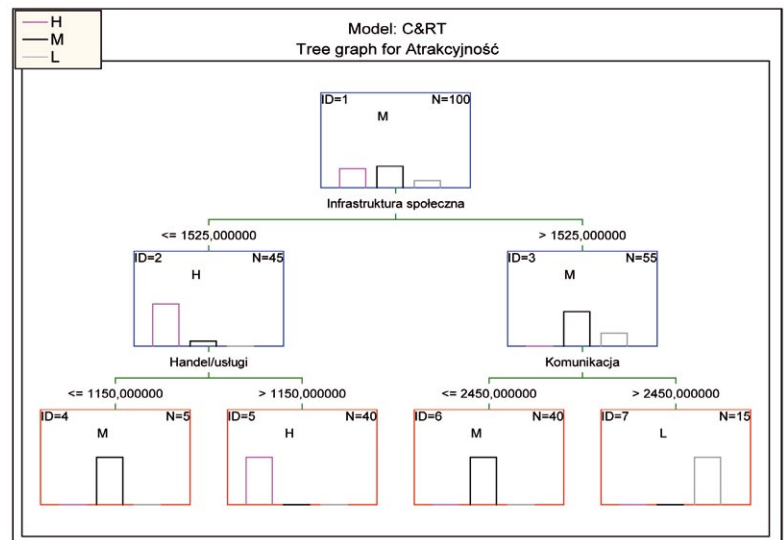


Rys. 1. Deklarowana znajomość zagadnień Big Data wśród kadry menadżerskiej średniego szczebla w przedsiębiorstwach sektora budowlanego (badania własne)



Rys. 2. Deklarowana dostępność zasobów danych w przedsiębiorstwach sektora budowlanego (badania własne)

Rys. 3. Przykład wyników analizy Data Mining w zakresie czynników lokalizacyjnych [6]



ograniczenia w stosowanej metodzie analizy. Zauważa się, że występowanie wraz z danymi tekstowymi wartości liczbowych w toku analizy Text Mining powoduje utratę kontekstu przez tę wartość i tym samym jej częściowe pominięcie. Pomimo tych niedogodności zastosowanie podejścia Text Mining daje szansę na relatywnie szybką i zautomatyzowaną analizę danych tekstowych.

Jedną z ostatnich aplikacji w analizie Big Data było jej wykorzystanie we wspomaganie zarządzania technicznego budynkami. Zasobami poddanymi analizie były dane tekstowe dotyczące stanu technicznego budynków na przestrzeni ostatnich 7–10 lat. Oprócz danych tekstowych w analizie uwzględniono dane ilościowe – księgowe, dotyczące ponoszonych przez zarządcę nakładów finansowych na utrzymanie w danym roku, a także inne dane charakterystyczne dla budynku (m.in. technologia, powierzchnia, liczba kondygnacji, wiek, struktura właścicielska). Prowadzona analiza początkowo nie była ukierunkowana i pozwoliła uzyskać interesujące zależności. Jedną z nich była zależność wskazująca, że przy udziale własnościowym spółki miejskiej wyższym aniżeli 38%, w budynkach starszych niż 68 lat, w technologii tradycyjnej stan techniczny był zauważalnie gorszy (stopień zużycia technicznego wyższy niż 61%), podczas gdy przy zmianie struktury własnościowej, gdzie liczba współwłaścicieli była mniejsza aniżeli 5, udział spółki miejskiej był mniejszy niż 17%, to stan techniczny był lepszy i poprzez wartość stopnia zużycia technicznego oznaczony został na poziomie niższym aniżeli 46%. Przedstawiona jako przykład zależność pozwala pozyskać istotne informacje z punktu widzenia zarządzania budynkiem, a w szczególności dane w zakresie kształtowania budżetu remontowego.

6. Podsumowanie

Obserwacja wykorzystania zasobów Big Data oraz analiz Data Mining w innych sektorach gospodarki i życia pozwala przypuszczać, że w sektorze budownictwa nastąpi znaczne przyspieszenie rozwoju aplikacji. W opinii autora rozwój cyfryzacji – ze szczególnym uwzględnieniem modelowania BIM, technik mobilnych przesyłu danych, dużej liczby gromadzonych danych spowoduje większe zainteresowanie zasobami Big Data i technikami ich analiz. Jest to o tyle stwierdzenie uprawnione, że już obecnie w sektorze budownictwa występują duże zasoby danych, których zautomatyzowana analiza może dostarczyć interesujących i cennych informacji.

BIBLIOGRAFIA

- [1] Berry M., Linoff G., Data mining techniques for marketing, sales and customer support. Wiley, New York, 1997
- [2] Bilal M., Oyedele L. O., Qadir J., Munir K., Ajayi S. O., Akinade O. O., Owolabi H. A., Alaka H. A., Pasha M., Big Data in the construction industry: A review of present status, opportunities, and future trends, *Advanced Engineering Informatics*, tom 30, 3/2016, str. 500–521, Doi: <http://dx.doi.org/10.1016/j.aei.2016.07.001>
- [3] Błażejewicz G., Rewolucja z Marketing Automation. Jak wykorzystać potencjał Big Data, Wydawnictwo Naukowe PWN, Warszawa, 2016
- [4] Gajzler M., Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry, *Technological and Economic Development of Economy*, 16(2)2010, str. 219–232, Doi: <https://doi.org/10.3846/tede.2010.14>
- [5] Gajzler M., Usefulness of mining methods in knowledge source analysis in the construction industry. *Archives of Civil Engineering*, tom 62, 1/2016, str. 127–142, Doi: <http://dx.doi.org/10.1515/ace-2015-0056>
- [6] Gajzler M., Zagadnienie wyboru lokalizacji z wykorzystaniem metody Data Mining, *Budownictwo i Inżynieria Środowiska*, tom 2, 3/2011, str. 253–261
- [7] Gajzler M., Protocols from periodic inspection of buildings in text mining analysis – advantages and problems of analysis, *MATEC Web of Conferences* 222/2018, str. 01004, Doi: <https://doi.org/10.1051/mateconf/201822201004>
- [8] Górecki J., Big Data as a Project Risk Management Tool. *Risk Management Treatise for Engineering Practitioner*, 2018, Doi: <http://dx.doi.org/10.5772/intechopen.79182>
- [9] Górecki J., Bizon-Górecka J., Michalkiewicz K., Execution of construction projects in the use of Big Data. *Scientific Review. Engineering and Environmental Science* 26(2)2017, str. 241–249, Doi: <http://dx.doi.org/10.22630/PNIKS.2017.26.2.23>
- [10] Ismail S. A., Bandi S., Maaz Z. N., Appraisal into the Potential Application of Big Data in the Construction Industry. *International Journal of Built Environment and Sustainability*, tom 5, 2/2018, str. 145–154, Doi: <http://dx.doi.org/10.11113/ijbes.v5.n2.274>
- [11] Qian Z., Yang X., Xu Z., Cai W., Research on Key Construction Technology of Building Engineering under the Background of Big Data, *IOP Conference Series: Earth and Environmental Science*. The 7th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2020) 14-15 November 2020, Busan, South Korea, Doi: <http://dx.doi.org/10.1088/1742-6596/1802/3/032003>
- [12] Tabakow M., Korczak J., Franczyk B., Big Data – definicje, wyzwania i technologie informatyczne, *Business Informatics* 1(31)2014
- [13] Yang C. T., Chen S. T., Liu J. C., Liu R. H., Chang C. L., On construction of an energy monitoring service using big data technology for the smart campus. *Cluster Computing*, 23/2020, str. 265–288, Doi: <https://link.springer.com/article/10.1007%2Fs10586-019-02921-5>
- [14] Zieliński J., *Inteligentne systemy w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa, 2000