CZASOPISMO INŻYNIERII LĄDOWEJ, ŚRODOWISKA I ARCHITEKTURY
JOURNAL OF CIVIL ENGINEERING, ENVIRONMENT AND ARCHITECTURE
JCEEA, t. XXXIII, z. 63 (1/I/16), styczeń-marzec 2016, s. 213-220

Janusz SZELKA[1]
Zbigniew WRONA[2]

MOŻLIWOŚCI WYKORZYSTANIA EKSPLORACYJNEJ ANALIZY DANYCH W PRZEDSIĘWZIĘCIACH INŻYNIERYJNYCH

Wszystkie kategorie procesów informacyjno-decyzyjnych, realizowanych w obszarze przedsięwzięć inżynieryjnych, wymagają gromadzenia i przetwarzania znacznych ilości danych. Systemy baz danych, eksploatowane w obszarze tych przedsięwzięć, wykorzystuje się niemal wyłącznie do bieżącego przetwarzania informacji. Ich wykorzystanie do celów analitycznych ogranicza się do analiz całkowicie sterowanych przez użytkownika (inżyniera). Natomiast, w wielu obszarach zarządzania, w przechowywanych zasobach danych dostrzega się ogromny potencjał analityczny i dokonuje się z powodzeniem ich zautomatyzowanej eksploracji, pozyskując w ten sposób nową wiedzę (odkrywając nietrywialne, nieznane wcześniej prawidłowości). Wydaje się, że nie ma przeszkód, by podobne działania realizować także w obszarze przedsięwzięć inżynieryjnych, odkrywając nowe klasyfikacje, asocjacje, czy identyfikując sekwencje zdarzeń. Zautomatyzowana eksploracja danych często okazuje się jedynym sposobem wyszukiwania prawidłowości w ogromnych zbiorach danych, których człowiek nie jest w stanie przeanalizować. Specyfika przedsięwzięć inżynieryjnych (znaczna złożoność, niejednorodność a często także niepowtarzalność sytuacji problemowych) narzuca przy tym określone ograniczenia na poszczególne etapy takiej analizy. W opracowaniu przybliżono uwarunkowania stosowania eksploracyjnej analizy danych w przedsięwzięciach inżynieryjnych, nakreślono zakres przedsięwzięć niezbędnych do wykonania w poszczególnych jej etapach oraz wskazano narzędzia umożliwiające programową realizację tego typu przedsięwzięć.

Słowa kluczowe: procesy analityczno-decyzyjne w przedsięwzięciach inżynieryjnych, eksploracyjna analiza danych, metody eksploracji danych w przedsięwzięciach inżynieryjnych

1. Wprowadzenie

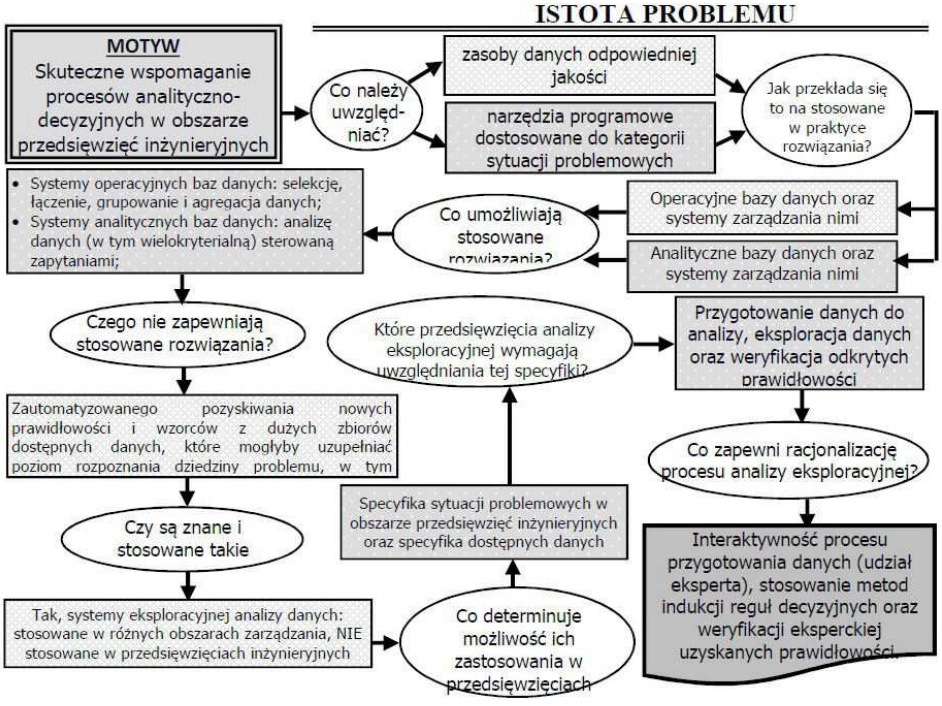
W szerokim spektrum działań inżynieryjnych coraz większego znaczenia nabiera problematyka automatyzacji procesów analityczno-decyzyjnych. Wykorzystanie w tym celu metod i narzędzi informatyki daje możliwość tworzenia coraz

[1] Autor do korespondencji/corresponding author: Janusz Szelka, Wyższa Szkoła Oficerska Wojsk Lądowych we Wrocławiu, Uniwersytet Zielonogórski, e-mail: jszelka@wso.wroc.pl
[2] Zbigniew Wrona, Wyższa Szkoła Zarządzania „Edukacja" we Wrocławiu, e-mail: z_wrona@wp.pl

bardziej złożonych i precyzyjnych opisów sytuacji problemowych, ale jednocześnie wymaga stosowania coraz sprawniejszych narzędzi sprzętowych i programowych do realizacji analiz. Dodatkowo, w szybkim tempie rozrastają się rozmiary wymaganych zasobów informacyjnych, choć nie przekłada się to na znaczące zwiększenie możliwości wykorzystania analitycznego potencjału tych danych.

Warto natomiast odnotować, że w niektórych obszarach zarządzania, w coraz większym stopniu dostrzega się znaczenie gromadzonych danych operacyjnych (ale także i historycznych) w odkrywaniu nowych prawidłowości, czy wzorców, co pociąga za sobą skuteczne próby zautomatyzowanej eksploracji tych danych. Analiza dostępnych zasobów informacyjnych w oparciu o ich eksplorację może okazać się użyteczna, a przy tym trudna do zastąpienia przy użyciu innych metod, także w obszarze przedsięwzięć inżynieryjnych. Warunkiem skutecznego wykorzystania eksploracyjnej analizy danych wydaje się odpowiednie przygotowanie zasobów informacyjnych, właściwy dobór metod i algorytmów eksploracji, uwzględniających specyfikę sytuacji problemowych w obszarze działań inżynieryjnych oraz prawidłowa interpretacja i weryfikacja rezultatów procesu eksploracyjnego.

Istotę rozpatrywanego problemu przedstawiono na rys. 1.



Rys. 1. Istota problemu

Fig. 1. The essence of the problem

2. Przedsięwzięcia analityczne w kontekście specyfiki procesów inżynieryjnych

Wyzwaniem w zakresie skutecznej realizacji procesów analityczno-decyzyjnych jest nie tylko dysponowanie odpowiednimi zasobami danych, lecz także możliwości ich analizowania, zdolność interpretacji i wyciągania użytecznych wniosków, które mogą prowadzić do racjonalnych decyzji.

Powszechnie wykorzystywanym narzędziem informatycznym, umożliwiającym gromadzenie danych (i informacji) w ustrukturyzowanej postaci są transakcyjne bazy danych. Dają one dostęp do szczegółowych danych operacyjnych (np. parametrów sprzętu przeprawowego, danych z systemu monitoringu, itp.), a obudowane o dodatkową powłokę programową, pozwalają na przetwarzanie tych danych w zakresie tzw. *OLTP* (ang. *OnLine Transactional Processes*), które obejmują selekcję, agregowanie, łączenie, czy grupowanie danych.

Realizacja złożonych przedsięwzięć analitycznych (np.: analizy porównawcze, predykcyjne, wielokryterialne) wymaga zastosowania specjalistycznej kategorii systemów baz danych, określanych mianem hurtowni danych. Są one zaopatrzone w rozbudowane mechanizmy wspomagające realizację wielowymiarowych analiz różnego typu (tzw. *OLAP* – ang. *OnLine Analytical Processes*). W obszarze przedsięwzięć inżynieryjnych hurtownie danych są stosowane rzadko, głównie z powodu bardzo dużych kosztów zakupu i eksploatacji, ale niektóre kategorie specjalistycznego oprogramowania, wspomagającego prace inżynierskie (typu CAD), zawierają moduły działające w oparciu o mechanizmy OLAP.

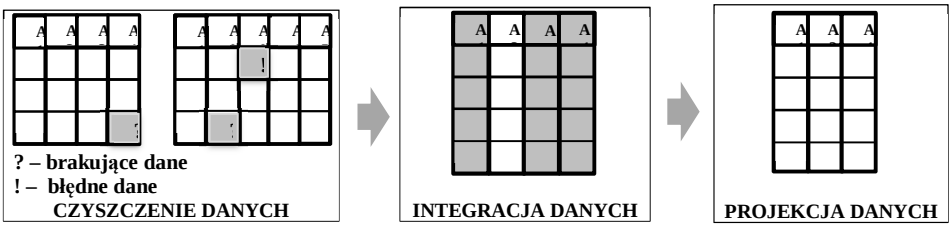
Warto jednak zauważyć, że posiadane zasoby danych operacyjnych (i ewentualnie analitycznych) można wykorzystać także do innego rodzaju analiz, w efekcie których mogą zostać odkryte niezidentyfikowane dotychczas regularności, asocjacje, czy sekwencje zdarzeń (nowa wiedza) – rys. 2.

Rys. 2. Analiza eksploracyjna w zestawieniu z procesami OLTP i OLAP

Fig. 2. Exploratory analysis in comparison with OLTP and OLAP

Poszukiwanie takich prawidłowości, przy użyciu określonych metod i algorytmów zgłębiania danych jest określane mianem eksploracji danych (ang. *Data Mining*) [1]. Natomiast termin eksploracyjnej analizy danych jest utożsamiany z procesem analizy posiadanych zbiorów danych, dokonywanej przy użyciu ich eksploracji. Proces ten powinien być nietrywialny, co oznacza, że musi wykraczać poza analizę, w której badane są tylko z góry określone zależności, co oferują proste aplikacje OLAP. Eksploracyjna analiza danych powinna umożliwiać poszukiwanie nowych, potencjalnie użytecznych i zrozumiałych struktur, wzorców lub modeli.

Eksploracja danych musi być poprzedzona przygotowaniem zasobów informacyjnych, przeznaczonych do analizy (rys. 3). Obejmuje ono m. in. proces czyszczenia danych, polegający na wyeliminowaniu obiektów (krotek) z ewidentnie błędnymi lub nieznanymi wartościami atrybutów, bądź też, zastąpieniu ich wartościami odpowiednio spreparowanymi. Przedsięwzięcia tego etapu można w pewnym stopniu automatyzować, np. zastępując brakujące wartości wartościami średnimi lub modalnymi (dla danych numerycznych). Wydaje się jednak, że specyfika sytuacji problemowych w obszarze przedsięwzięć inżynieryjnych (sytuacje wielokrotnie niepowtarzalne oraz występowanie znacznej ilości danych nienumerycznych w bazach danych, skłania do wniosku, że przygotowanie danych w tym zakresie powinno być nadzorowane przez eksperta (inżyniera).



Rys. 3. Proces przygotowania danych do analizy eksploracyjnej

Fig. 3. Process of data preparation for exploratory analysis

Dane wykorzystywane w procesie eksploracji mogą znajdować się w wielu zbiorach danych, posiadających różne struktury lub formaty, które należy ujednolicić (w zakresie nazw atrybutów, dziedzin atrybutów, a w niektórych przypadkach także przeprowadzić ich standaryzację, czy kategoryzację). Następnie należy wydzielić te atrybuty, które mają zostać poddane procesowi analizy. Do realizacji tego zadania można wykorzystać wiedzę eksperta bądź, w niektórych przypadkach, narzędzia dokonujące zautomatyzowanej oceny ważności atrybutów, np. w oparciu o algorytm *Minimum Description Length*, dostępny, między innymi, w aplikacji *Oracle Data Miner* (Oracle Corporation).

Zasadnicze przedsięwzięcie analizy eksploracyjnej, określane mianem eksploracji danych, może być realizowane przy użyciu różnorodnych metod, technik i narzędzi. Dobór metod zależy w znacznej mierze od specyfiki sytuacji

problemowej, celu eksploracji, czy rodzaju eksplorowanych danych. Uwzględniając specyfikę sytuacji problemowych w obszarze przedsięwzięć inżynieryjnych, za metody o dużej, potencjalnej użyteczności można uznać metody klasyfikacji obiektów, generujące tzw. reguły klasyfikacyjne, odkrywane najczęściej w oparciu o zbiór danych treningowych, a następnie wykorzystywane dla nowych sytuacji oraz metody okrywania asocjacji – obejmujące odkrywanie zależności, opisywanych następnie w zbiorach reguł asocjacyjnych.

Do odkrywania reguł mogą być stosowane różne algorytmy, różniące się m. in. wydajnością, czy sposobem przygotowania danych wejściowych. Na przykład, do odkrywania reguł asocjacyjnych, można wykorzystać algorytm *Apriori*, a do odkrywania klasyfikacji – *SVM* [1].

Prawidłowości odkryte w procesie eksploracji danych należy ocenić w zakresie ich wiarygodności oraz użyteczności. Pierwszy z aspektów może być częściowo zautomatyzowany, dzięki określeniu wartości kilku wskaźników jakości reguł, wypracowanych w rezultacie eksploracyjnej analizy danych. Dwa najczęściej stosowane wskaźniki, to *wsparcie reguły* (ang. *Support*) - liczba krotek relacji (lub grup krotek) potwierdzających odkrytą regułę oraz *zaufanie reguły* (ang. *Confidence*) – stosunek liczby krotek relacji spełniających regułę do liczby krotek relacji, dla których jest spełniona część warunkowa reguły [2].

W praktyce zarządzania, przedsięwzięcia eksploracyjnej analizy danych są realizowane przez specjalistyczne systemy Data Mining, wspomagające zarówno procesy przygotowania danych, jak i ich eksploracji. Można do nich zaliczyć m. in.: *Statistica Data Miner* (Statsoft), czy *Oracle Data Miner* (Oracle).

3. Wykorzystanie metod i narzędzi eksploracyjnej analizy danych w przedsięwzięciach inżynieryjnych

Niezależnie od jakości danych, którymi dysponujemy w procesie analizy eksploracyjnej, należy założyć, że ich struktura będzie miała postać tabelaryczną (dane gromadzone np. w arkuszach kalkulacyjnych) lub relacyjną (operacyjne bazy danych). Przykładową tabelę, zawierającą dane treningowe, dotyczące kolejnych wpisów obejmujących parametry przeszkody wodnej oraz odpowiadających im, dobranych konstrukcji podpór zaprezentowano w tab. 1. W efekcie procesu czyszczenia danych (usuwanie lub zastępowanie przez eksperta brakujących lub niewłaściwych danych) oraz integracji danych (łączenia analogicznych tabel z różnych źródeł i ujednolicenia atrybutów) uzyskano postać tabeli z danymi źródłowymi. Następnie, w oparciu o ujęcie modelowe dla mostów składanych [3], poddano procesowi kategoryzacji atrybuty ilościowe: *Szerokość przeszkody wodnej* (SPW) [m], *Głębokość przeszkody wodnej* (GPW) [m], *Głębokość koryta rzeki* (GKR) [m], *Szybkość prądu rzeki* (SPR) [m/s] oraz *Przewyższenie brzegów* (PB) [m]. Przykładowo, dla atrybutów: SPW oraz GKR, przyjęto następujące kategorie:

(SPW): *Sp1* <= 40; 40 < *Sp2* <=100; 100 <*Sp3* <= 300; *Sp4* > 300

(SPR): *Vp1* <= 0,5; 0,5 < *Vp2* <= 1,5; 1,5 < *Vp3* <=2,5; *Vp4* >2,5

Tablica 1. Proces eksploracji reguł klasyfikacyjnych – przygotowanie danych
Table 1. Process of exploration of classification rules – data preparation

NR PRZEPRAWY	SPW	GPW	GKR	SPR	RODZAJ GRUNTU DNA (RGD)	RODZAJ GRUNTU BRZEGÓW (RGB)	PB	RODZAJ PODPORY (RP)	NUMER NADBUDOWY (NN)	LICZBA PODPÓR POŚREDNICH (LP)	WYSOKOŚĆ RUSZTU (WR)	WYSOKOŚĆ NADBUDOWY (WN)
1	170 / Sp3	4,5 / Hw3	12 / Hk3	1,2 / Vp2	spoisty	drobno-ziarnisty	1 / Pb1	SPS-69 B	P246	5	5,5	7,5
2	120 / Sp3	3,5 / Hw2	7,5 / Hk2	0,8 / Vp2	gruboziarnisty	spoisty	1,5 / Pb2	SPS-69 B	P123	3	4,5	4,5
3	80 / Sp2	2,5 / Hw2	5 / Hk1	1,5 / Vp2	gliniasty	spoisty	2,5 / Pb2	SPS-69 B	P83	2	3.0	4
.....												
40	168 / Sp3	4,8 / Hw3	11,5 / Hk3	1,0 / Vp2	spoisty	drobno-ziarnisty	0,8 / Pb1	SPS-69 B	P246	5	6,0	7,5

W wyniku indukcji reguł klasyfikacyjnych, wygenerowanych zostanie szereg reguł decyzyjnych. Weryfikacja wstępnej wiarygodności odkrytych reguł, dokonana w oparciu o zależność *sup*(Ri) >= *minsup* (przy założeniu, że *minsup* zostało określone przez eksperta na poziomie 20%) pozwala zakwalifikować do dalszej weryfikacji reguły R1 oraz R2, natomiast sup(R3) < minsup – tab. 2.

Następnie, dla zakwalifikowanych wstępnie reguł, w oparciu o zależność: *con*(Ri) >= *mincon*, gdzie próg zaufania reguły został określony przez eksperta na poziomie 25%, uzyskuje się zestaw wiarygodnych reguł (w rozpatrywanym przypadku reguły: R1 oraz R2), które należy poddać ocenie eksperta w zakresie ich użyteczności.

Z kolei, odkrywanie asocjacji w zbiorze danych, zaprezentowanym na rys. 4, pozwala określić zależności pomiędzy uszkodzeniami, obserwowanymi w przypadku obiektów mostowych. Wykorzystując algorytm *Apriori* można uzyskać m. in. zależności R1-R7, opisujące sytuacje, w których wystąpienie określonego uszkodzenia często wiąże się z innymi, ściśle określonymi uszkodzeniami tego obiektu oraz przypisane poszczególnym regułom wartości współczynnika wsparcia (sup):

Tablica 2. Przykłady odkrytych reguł klasyfikacyjnych
Table 2. Examples of discovered classification rules

(R1)	**IF** SPW='Sp3' AND GPW='Hw3' AND GKR = 'Hk3' AND SPR='Vp2' AND RGD='spoisty' AND RGB='drobnoziarnisty' AND PB='Pb1' AND RP='SPS-69 B' **THEN** NN='P246' AND LP=5 AND WN=7,5
sup(R1)=20%, con(R1)=30%	
(R2)	**IF** SPW='Sp3' AND GPW='Hw2' AND GKR = 'Hk2' AND SPR='Vp2' AND RGD='gruboziarnisty' AND RGB='spoisty' AND PB='Pb2' AND RP='SPS-69 B' **THEN** NN='P83' AND LP=2 AND WR=3 AND WN=4
sup(R2)=20%, con(R2)=25%	
(R3)	**IF** SPW='Sp2' AND GPW='Hw2' AND GKR = 'Hk1' AND SPR='Vp2' AND RGD='gliniasty' AND RGB='drobnoziarnisty' AND PB='Pb2' AND RP='SPS-69 B' **THEN** NN='P123' AND LP=3 AND WR=3 AND WN=7,5
sup(R3)=10% < minsup → (R3) - odrzucona	

NR OBSER-WACJI	NR OBIEKTU MOSTOW.	RODZAJ ZNISZCZENIA
1	1	podmycie podpór/ przewrócenie
1	1	załamanie się przęseł / zniszczenie
2	2	podmycie pod-pór/przewrócenie
2	2	załamanie się przęseł / zniszczenie
2	2	przesunięcie nurtu rzeki
3	3	podmycie pod-pór/przewrócenie
3	3	uszkodzenia nawierzchni
3	3	przesunięcie nurtu rzeki
4	4	podmycie podpór/ przewrócenie
4	4	załamanie się przęseł / zniszczenie

NR OBSER-WACJI	Podmycie podpór/przewrócenie (Z1)	Załamanie się przęseł/zniszczenie (Z2)	Przesunięcie nurtu rzeki (Z3)	Uszkodzenie nawierzchni (Z4)
1	1	1	0	0
2	1	1	1	0
3	1	0	1	1
4	1	1	0	0

Binarna tablica obserwacji

	sup	*con*
(R1) IF (Z1) THEN (Z2)	75%	75%
(R2) IF (Z1) AND (Z2) THEN (Z3)	25%	33%
(R3) IF (Z1) THEN (Z2) AND (Z3)	25%	25%
(R4) IF (Z1) THEN (Z3)	50%	50%
(R5) IF (Z1) THEN (Z4)	25%	25%
(R6) IF (Z1) AND (Z3) THEN (Z4)	25%	50%
(R7) IF (Z1) THEN (Z3) AND (Z4)	25%	25%

Rys. 4. Proces odkrywania reguł asocjacyjnych

Fig. 4. Process of exploration of association rules

Określenie przez eksperta minimalnego wsparcia dla reguły *minsup* na poziomie 30% pozwala zakwalifikować do wstępnie przyjętego zbioru reguł, reguły R1 oraz R4, a przy ustaleniu: *mincon* = 50%, zarówno R1, jak i R4 zostaną uznane za wiarygodne. Użyteczność odkrytych reguł powinien zweryfikować ekspert.

W oparciu o podany przykład można również zrealizować analizę eksploracyjną wzorców sekwencji, rozszerzając analizę asocjacji o kolejność (sekwencję) uszkodzeń obiektu mostowego.

4. Podsumowanie

W wielu obszarach zarządzania dostępne zbiory danych są traktowane obecnie nie tylko jako zasoby, stanowiące zasilenie informacyjne procesów analityczno-decyzyjnych. Postrzega się je również jako źródło nowych, potencjalnie użytecznych prawidłowości i wzorców – źródło cennej często wiedzy, odkrywanej w zautomatyzowany sposób. Podobne własności można przypisać dużym zbiorom danych operacyjnych i analitycznych, eksploatowanym w obszarze przedsięwzięć inżynieryjnych. Wykorzystywane dotychczas jedynie w zakresie analiz sterowanych całkowicie przez użytkownika, nie pozwalały na odkrywanie analitycznego potencjału zgromadzonych danych. Poddanie ich procesowi analizy eksploracyjnej umożliwia zautomatyzowane przeszukiwanie tych zasobów pod kątem odkrywania nowej wiedzy. Jej wiarygodność oraz użyteczność jest uwarunkowana jedynie narzuceniem określonych ograniczeń na poszczególne etapy procesu analizy oraz na zastosowane metody eksploracji danych. Należy założyć, że analiza eksploracyjna, dokonywana w obszarze przedsięwzięć inżynieryjnych

powinna być procesem interaktywnym i iteracyjnym. Udział eksperta wydaje się nieodzowny już na etapie przygotowania danych, a w szczególności, w ramach przedsięwzięć czyszczenia danych (uzupełniania brakujących danych) oraz projekcji danych (wyboru atrybutów istotnych w procesie eksploracji). Podobna uwaga dotyczy etapu weryfikacji otrzymanych prawidłowości, a w szczególności oceny ich przydatności (użyteczności). Po otrzymaniu i przeanalizowaniu uzyskanych wzorców, ekspert ma możliwość skorygowania parametrów zadania przez zawężenie/rozszerzenie zbioru eksplorowanych danych lub zawężenie/rozszerzenie zbioru poszukiwanych wzorców – po czym może nastąpić przejście do kolejnej iteracji procesu analizy danych.

Literatura

[1] Larose D. T. Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych. PWN, Warszawa 2013.

[2] Stefanowski J. Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy. Rozprawa habilitacyjna. Wydaw. Politechniki Poznańskiej, Poznań 2001.

[3] Szelka J., Obiektowy zapis wiedzy w systemach eksperckich wspomagających budowę mostów wojskowych, WAT, Warszawa 1999.

THE POSSIBILITY OF USING EXPLORATORY DATA ANALYSIS IN ENGINEERING PROJECTS

S u m m a r y

In all categories of information-based decision-making processes implemented in the area of engineering, a significant amount of data must be gathered and processed. Parameters of engineering equipment, as well as data gathered, inter alia, during analysis, design and construction of engineering objects or systems for monitoring engineering structures are stored mainly in operational databases. Database systems utilized in the area of engineering are used almost exclusively for ongoing information processing. Their use for analytical purposes is limited to analyses entirely directed by the user (engineer). On the other hand, in many areas of management, pools of stored data are valued for their immense analytical potential, and their automated exploration is successfully conducted, yielding new knowledge (bringing out extraordinary, hitherto unknown regularities). There is no reason to believe that such activities would not be feasible also in the area of engineering, where they would produce discoveries of new classifications, associations, or identification of sequences of events. Automated exploration of data often turns out to be the only way of looking for regularities in pools of data which are too large for a human being to analyze. The character of engineering projects (their high complexity, heterogeneity, and often the uniqueness of the problem situation) imposes specific restrictions on each phase of such analysis. This study explains conditions for use of exploratory data analysis in engineering projects, delineates the scope of activities which have to be undertaken on its consecutive stages, and presents tools enabling programmatic completion of such projects.

Keywords: analytical decision-making processes in engineering projects, exploratory data analysis, methods of data exploration in engineering projects

Przesłano do redakcji: 07.06.2016 r.
Przyjęto do druku: 30.06.2016 r.

DOI: 10.7862/rb.2016.25