ANALIZA I OCENA WYBRANYCH MODELI EKSPLORACJI DANYCH

Celina M. OLSZAK, Kamila BARTUŚ

Streszczenie: W artykule przedstawiono wybrane problemy związane z eksploracją danych. Zaprezentowano i opisano istniejące modele procesowe eksploracji danych oraz dokonano ich oceny. W dalszej kolejności przedstawiono własną propozycję modelu procesowego eksploracji danych oraz jego weryfikację.

Słowa kluczowe: eksploracja danych, CRISP-DM, SEMMA, DMAIC, VCofDM.

1. Pragmatyka i cele eksploracji danych

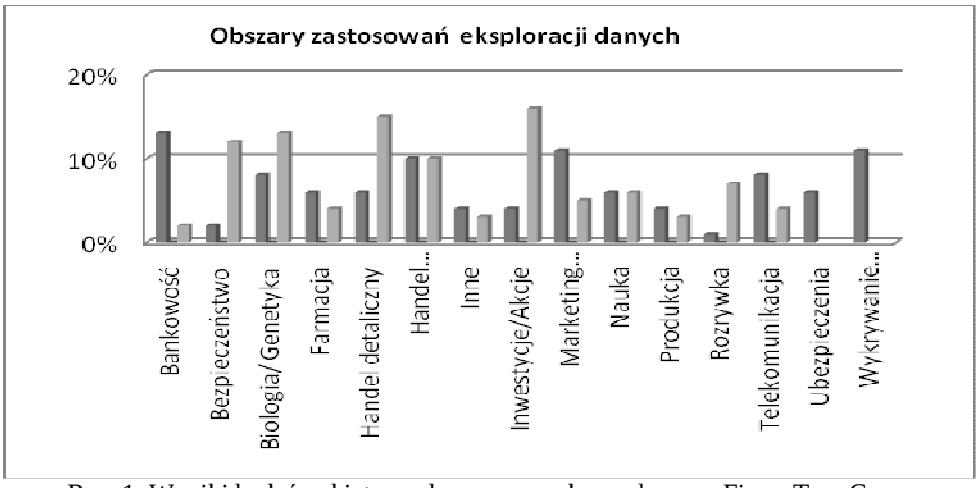
Rozpoczynając rozważania na temat eksploracji danych, w pierwszej kolejności, należy zwrócić uwagę, czym eksploracja danych nie jest. Uważa się, że nie jest ona [6, 10, 9]:

- odkrywaniem wiedzy - eksploracja danych jest często osadzana w szerokim kontekście odkrywania wiedzy z danych (Knowledge Discovery in Database - KDD) i liczni autorzy traktują te pojęcia zamiennie, bądź też jako synonimy ograniczając tym samym proces KDD do jego analitycznej części,
- klasycznym narzędziem generującym analizy i sprawozdania – w oparciu o narzędzia analityczne (np. OLAP) dokonuje się przede wszystkim weryfikacji hipotez zaproponowanych przez ekspertów/analityków, natomiast eksploracja danych umożliwia odkrywanie zasad, reguł oraz hipotez,
- procesem automatycznym, który nie wymaga udziału ludzkiego nadzorowania – eksploracja danych jest w dużym stopniu uzależniona od wiedzy eksperta, który określa problem biznesowy, przekłada go na problem eksploracji danych, dobiera metody i techniki eksploracji danych, a następnie ocenia uzyskane wyniki,
- procesem, który odszuka przyczyny problemów biznesowych – eksploracja danych umożliwia odnajdywanie wzorców czy trendów pewnych zjawisk, ale tylko ekspert jest w stanie zidentyfikować ich przyczyny,
- łatwym i szybkim procesem – eksploracja danych jest złożonym procesem, który wymaga od ekspertów interdyscyplinarnej wiedzy oraz szerokiego wachlarza doświadczeń.

Najczęściej przytaczaną definicją, jaką można spotkać w literaturze przedmiotu jest definicja mówiąc, iż eksploracja danych to proces odkrywania istotnych zależności, korelacji, wzorców i tendencji, poprzez przesiewanie dużych ilości danych przechowywanych w repozytoriach za pomocą technik rozpoznania wzorców oraz technik statystycznych i matematycznych.

Eksploracja danych zdobywa coraz większą popularność jako sposób wydobywania ukrytej wiedzy ze zbiorów danych. Podyktowane jest to faktem rosnącego popytu na wiedzę oraz lawinowego zwiększania się ilość danych, opisujących zarówno organizacje, jak i ich otoczenie [7]. Eksploracja danych znajduje swoje zastosowanie w wielu dziedzinach gospodarki. Badania empiryczne, przeprowadzone w latach 2004-2006 przez

firmy Two Crows oraz KDnuggets, dowodzą, iż jest ona najczęściej wykorzystywana przez takie sektory jak: bankowość, handel elektroniczny, marketing ukierunkowany, bezpośredni oraz krzyżowy, telekomunikację, medycynę (Rys. 1).



Rys. 1. Wyniki badań ankietowych przeprowadzonych przez Firmę Two Crows Corporation oraz KDnuggets [13, 15]

Wyniki badań potwierdzają, że duże korzyści ze stosowania eksploracji danych mogą osiągnąć organizacje działające na dowolnym obszarze rynku. Dotyczy to, zarówno organizacji, które do tej pory nie wykorzystywały żadnych rozwiązań umożliwiających analizę danych, jak również tych, które już wykorzystują różne rozwiązania informatyczne.

2. Charakterystyka i ocena modeli eksploracji danych

Jak już wcześniej podkreślono, eksploracja danych oznacza proces, dzięki któremu możliwe staje się odnalezienie zbioru cennych informacji spośród ogromnej ilości ukrytych danych [5]. Słowo proces jest niezwykle istotne w tym określeniu. W wielu środowiskach istnieje przekonanie, że na eksplorację danych składa się tylko wybór i zastosowanie narzędzi, które będą w stanie sprostać postawionemu problemowi i automatycznie dostarczą gotowych odpowiedzi [1]. Jest to dosyć wąski punkt widzenia, ponieważ eksploracja danych nie jest zbiorem odizolowanych narzędzi, których sama implementacja usprawni rozwiązanie problemów decyzyjnych organizacji. Udana eksploracja danych, bowiem wymaga systemowego, modelowego przejścia przez sekwencję etapów. W związku z tym, pojawiła się wyraźna potrzeba tworzenia modeli eksploracji danych, które pomagałyby organizacjom wprowadzić w praktykę własne projekty i inicjatywy. Umożliwiałyby one także otrzymywanie lepszych rezultatów eksploracji i promowałyby najlepsze praktyki biznesowe.

Do 1996 roku nie istniało żadne powszechnie akceptowane podejście do tworzenia eksploracji danych. Obecnie proponowane są różne modele, do których można zaliczyć: CRISP-DM, SEMMA, DMAIC, VCofDM.

Twórcami modelu CRISP-DM (Cross-Industry Standard Process for Data Mining) są NCR Systems Engineering Copenhagen (Dania), Daimler-Chrysler (Niemcy), SPSS/Integraf Solutions Ltd. (Anglia) oraz OHRA Verzekeringen Bank Group B.V. (Holandia) [11].

Według ich opinii na model eksploracji danych składa się sześć etapów:

- zrozumienie uwarunkowań biznesowych –odpowiada za zrozumienie uwarunkowań biznesowych i przełożenie ich na problem eksploracji danych. Sformułowany problemu eksploracji danych, zdaniem twórców tegoż modelu, nie powinien być wyrażany szerokimi, ogólnymi terminami. Raczej zalecane jest, jeśli to możliwe, rozbicie ich na mniejsze, bardziej szczegółowe;
- zrozumienie danych – na tym etapie przeprowadza się wstępną analizę danych mającą na celu zaznajomienie się z danymi. Szczególny nacisk położono tutaj na dokładne zweryfikowanie jakości danych oraz ich kompletności. Bowiem na tym etapie sprawdza się reguła GIGO (wprowadzisz błędne dane, uzyskasz błędne wyniki). Brakujące i błędne dane mogą stanowić groźny problem podczas budowy modelu eksploracji danych;
- przygotowanie danych – etap ten jest jednym z najbardziej pracochłonnych etapów całego procesu eksploracji danych. Wiąże się on z przygotowaniem ze wstępnych, surowych danych ostatecznego zbioru danych, który będzie przedmiotem dalszego drążenia. Czynności jakie należy wykonać na tym etapie to przede wszystkim: eliminacja nieistotnych lub niepotrzebnych danych, które następnie zostają poddane weryfikacji poprawności i oczyszczaniu; wybór przypadków i zmiennych, które będą analizowane i które są odpowiednie do analizy;
- modelowanie – obejmuje wybór technik, które zostaną użyte do stworzenia modelu eksploracji danych. Istnieje kilka technik, które stosuje się do rozwiązania tego samego problemu eksploracji danych, co związane jest z określonymi wymaganiami np. formy danych. W takim wypadku, sugeruje się, konieczność powrotu do etapu przygotowania danych;
- ewaluacja – etap ten obejmuje przede wszystkim: ocenę modelu lub kilku modeli, otrzymanych z etapu modelowania; ustalenie, czy model spełnia wszystkie założenia ustalone w pierwszym etapie; sprawdzenie, czy jakieś cele biznesowe nie zostały pominięte; podjęcie decyzji co do wykorzystania eksploracji danych. Przed przejściem do ostatecznego wdrożenia modelu istotne jest dokładne przetestowanie i ponowne przejrzenie jego konstrukcji. Na tym etapie niezwykle ważne jest określenie, czy istnieje jakikolwiek problem biznesowy, który nie został wystarczająco należycie rozważony. Jeśli takowy istnieje, należy przejść do etapu pierwszego i powtórnie rozpocząć cały cykl eksploracji danych [8];
- wdrożenie – podczas tego etapu należy dokonać syntezy wiedzy zdobytej w trakcie trwania całego procesu eksploracji danych oraz przedstawić ją użytkownikowi w wizualny sposób. Kluczowymi krokami podczas tego etapu jest monitorowanie podjętych działań, stworzenie raportu końcowego oraz ponowna analiza projektu.

Natomiast model procesowy SEMMA (Sample, Explore, Modify, Model, Assess), został zaprojektowany przez SAS Institute.

Składa się on z pięciu etapów, w skład których wchodzą etapy tj. [14]:

- próbkowanie – polega na wykorzystaniu tylko części danych (przed wprowadzeniem ich większej ilości). Jeżeli podstawowym celem eksploracji jest odkrycie ogólnych wzorców czy też tendencji w danych, to zastosowanie

próbkowania umożliwia osiągnięcie takich rezultatów w stosunkowo krótkim czasie;

- eksplorowanie – etap ten odpowiedzialny jest za wnikliwe przeszukiwanie i eksplorowanie danych w celu ich dogłębnego poznania. Możliwe jest to dzięki znacznemu rozwojowi metod i technik eksploracji danych, do których można zaliczyć m.in.: odkrywanie asocjacji, klasyfikację, analizę skupień, sieci neuronowe, drzewa decyzyjne oraz algorytmy genetyczne;
- manipulacja – po etapie eksplorowania danych czasami zachodzi potrzeba modyfikacji lub manipulacji danymi. Twórcy modelu zaznaczają, iż dynamiczny charakter eksploracji danych lub większe uaktualnienie baz danych na potrzeby eksploracji powinny być możliwe na wszystkich etapach tego procesu;
- modelowanie – na tym etapie następuje wybór technik modelowania, do których autorzy modelu zaliczają przede wszystkim drzewa decyzyjne, modele szeregów czasowych oraz wywodzące się ze sztucznej inteligencji np. sieci neuronowe;
- ocena – jest ostatnim etapem modelu SEMMA, który odpowiada za porównanie i ocenę wcześniej skonstruowanych modeli eksploracji danych;

Kolejnym wyżej wymienionym modelem jest model DMAIC (Define, Measure, Analyze, Improve, Control), który oparty jest na strategii Six Sigma i został zainicjowany przez grupę inżynierów Instytutu Motoroli.

Początkowo strategia ta skupiona była na stworzeniu tzw. projektu zwycięskiego wyrobu [2]. Teraz koncentruje się na eliminacji defektów, strat i problemów z jakością we wszystkich dziedzinach biznesu [4].

Model ten składa się z pięciu etapów, które biorą swój początek z tradycji ulepszenia jakości. Należą do nich:

- definiowanie – etap ten obejmuje określenie celów oraz identyfikację problemów biznesowych, których rozwiązanie pomoże w podwyższeniu poziomu Sigma,
- pomiar – na tym poziomie gromadzone zostają informacje dotyczące aktualnego stanu procesu w celu ustalenia poziomu odniesienia oraz rozpoznania skali problemu,
- analiza – w fazie tej zidentyfikowane zostają krytyczne przyczyny problemów dotyczących jakości oraz uzasadnienie ich wpływu na proces,
- usprawnienie – etap ten odpowiada za wprowadzenie rozwiązań, które pomogą w usunięciu analizowanych wcześniej problemów,
- kontrola – podczas tego etapu uzyskane wyniki zostają zweryfikowane i poddane obserwacji.

Ostatnim z wymienionych modeli jest model VCofDM (Virtuous Cycle of Data Mining), zaprojektowany on został przez wybitnych specjalistów z dziedziny eksploracji danych M. J. A. Berrego i G. Linoffa. Składa się z czterech etapów, powtarzanych cyklicznie, które uzupełnione są zbiorem dodatkowych wskazówek. Należą do nich [3]:

- zidentyfikowanie problemów biznesowych – zdaniem autorów na tym etapie należy identyfikować identyfikować te kwestie działalności gospodarczej, które mogą zostać wydatnie wsparte przez wykorzystanie wiedzy pozyskanej dzięki eksploracji danych. W modelu założono, że identyfikacja problemu biznesowego jest domeną osoby, która jest związana z konkretnym obszarem działalności gospodarczej. To do jej kompetencji należy sformułowanie właściwego celu biznesowego, określenie oczekiwań dotyczących uzyskanych odpowiedzi oraz zaplanowanie przyszłego ich wykorzystania. Celem pełnego zrozumienia przez

analityka eksploracji danych problemu biznesowego i późniejsze przełożenie go na przedsięwzięcie eksploracji danych, konieczne jest nawiązanie obustronnej współpracy między nim a osobą odpowiedzialną za podejmowanie decyzji w organizacji. Dzięki temu, możliwe staje się właściwe rozpoznanie i uwzględnienie potrzeb biznesowych w procesie eksploracji danych. Autorzy sugerują także, że podsumowanie pierwszego etapu powinno zawierać propozycje miar i sposobów oceny skuteczności podjętych działań oraz jakości wyników uzyskanych z eksploracji danych;

- przekształcenie danych w informacje – według autorów tego modelu, etap ten obejmuje kilka kluczowych procesów związanych z danymi. W skład nich wchodzi: zidentyfikowanie i pozyskanie danych odzwierciedlających dany problem biznesowy, przygotowanie ich zgodnie z wymogami narzędzi eksploracji danych oraz wybór odpowiedniej techniki modelującej. Prace analityka danych związane są głównie z pozyskaniem i przygotowaniem danych pochodzących z różnych źródeł, ujednoliceniem ich struktury, dodaniu dodatkowych pól, przygotowaniem zbioru danych do eksploracji, zaprojektowaniem modelu eksploracji danych oraz sprawdzeniem jakości otrzymanych wyników;
- podjęcie działań – autorzy VCoDM, już na pierwszym etapie modelu wskazują konieczność wykorzystania uzyskanych wyników w celu poprawy jakości podjętych działań. Jednak podjęcie decyzji na podstawie uzyskanych danych, to nie jedyne działania jakie należy podjąć. Proponuje się również realizację prac zmierzających do zapamiętania uzyskanych wyników, upowszechnienia zgromadzonej wiedzy, cyklicznego przewidywania (np. comiesięcznej predykcji wykorzystania wybranego produktu/usługi), poprawy jakości zgromadzonych danych oraz przeprowadzenia nowego projektu eksploracji danych związanego pośrednio z odkrytymi zależnościami;
- mierzenie i ocena wyników – Podczas oceny uzyskanych wyników należy porównać obecne wartości mierników z otrzymanymi z eksploracji danych. Zaznacza się, że szacowane wartości miar powinny odpowiadać prognozom błędu, jakimi cechują się otrzymane wyniki. Dlatego zaleca się sprawdzenie, jak te przewidywania korelują z obecną sytuacją oraz jakie były/są efekty przeprowadzonych działań.

Jak wynika z powyższych rozważań, zarówno w nauce jak i w praktyce podejmowano próby opracowania różnych modeli, które byłyby pomocnym scenariuszem pokazującym, w jaki sposób pozyskać, przygotować i przeanalizować dane, interpretować uzyskane wyniki oraz oceniać rezultaty tych działań. Zaprezentowane wyżej modele, pomimo wspólnego celu jakim jest wykorzystanie eksploracji danych do odkrycia nowej wiedzy, reprezentują różne podejścia do tej kwestii. Względnie duża dowolność definiowania zagadnień oraz etapów, jakie powinny zostać ujęte w procesie eksploracji danych, zrodziła konieczność ich analizy oraz oceny.

Wśród istotnych zalet analizowanych modeli, w przypadku CRISP-DM wymienić można: wskazanie dwóch podmiotów (klienta i analityka danych), które realizują projekt. W przypadku modelu SEMMA przywiązuje się dużą wagę do szczegółowego opisu poszczególnych kroków eksploracji danych. W modelu VCoDM wskazuje się na konieczność stałej współpracy, (na każdym etapie) między analitykiem danych a ekspertem dziedzinowym.

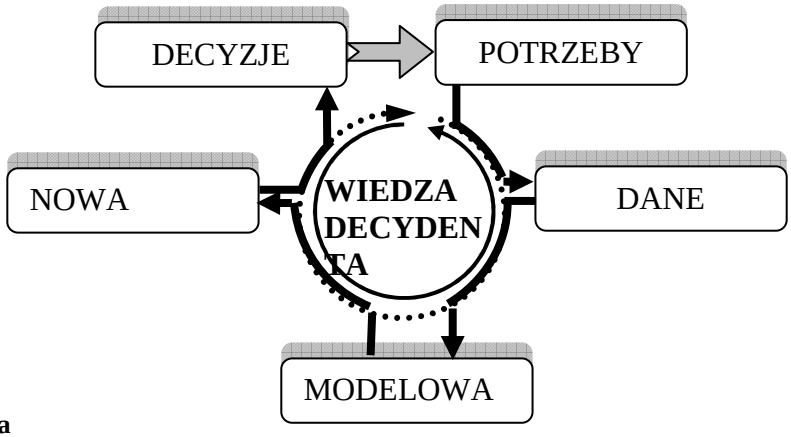
Wadą analizowanych modeli, jest pominięcie w nich miejsca i roli podmiotu, który steruje tym procesem. Zbyt mały nacisk kładzie się na archiwizację wiedzy wygenerowanej

podczas procesu eksploracji danych. Dotyczy ona zarówno odkrytych zależność i trendów pomiędzy danymi, jak również sposobu pozyskania danych, wykorzystania wybranych narzędzi informatycznych (ze wskazaniem konkretnych metod i technik), zasad budowy modeli oraz interpretacji uzyskanych wyników.

Wadą modelu CRISP-DM jest nadanie pełnej autonomii (wręcz odrębności) podmiotowi realizującemu proces eksploracji danych, co może prowadzić do pominięcia w nim ważnych aspektów dotyczących działalności gospodarczej. Minusem modelu SEMMA jest przede wszystkim nacechowanie go technicznymi aspektami. Natomiast modelowi DMAIC zarzuca się zbyt ogólny opis poszczególnych etapów eksploracji danych.

3. Propozycja modelu eksploracji danych oraz korzyści z jego weryfikacji

Przeprowadzona analiza modeli eksploracji danych pozwoliła stwierdzić, że zbyt mało uwagi poświęca się w nich wspomaganiu pracy decydenta. Dotychczasowe propozycje wyraźnie skierowane są na analityków lub informatyków. Decydent bez ich pomocy nie jest w stanie pracować z takimi modelami. W tej sytuacji pojawiła się potrzeba stworzenia modelu, który umożliwiałby decydentowi w elastyczny i intuicyjny sposób pracować na danych i na każdym etapie otrzymywać wsparcie. Proponowany model PDMWD (Rys. 2.) powstał w wyniku przeprowadzonego eksperymentu badawczego, łączącego wiedzę teoretyczną oraz doświadczenie autorów niniejszego artykułu w zakresie wykorzystywania modeli CRISP-DM, SEMMA, DMAIC oraz VCofDM.



Legenda

Strzałki reprezentują główne zależności między etapami. Okrąg wewnętrzny jest symbolem cyklicznego charakteru procesu eksploracji danych, którego przebieg, jakość oraz kierunek zależne są od preferencji i wiedzy decydenta.

Rys. 2. Model procesowy PDMWD

Nazwa modelu PDMWD pochodzi od akronimu opisującego główne jego etapy. Są to:

- potrzeby – etap ten związany jest z identyfikacją potrzeb informacyjno-decyzyjnych. Należy zaznaczyć, iż zasadnicze znaczenie dla zidentyfikowania problemu ma właściwa komunikacja nawiązana między analitykiem danych oraz decydentem. Sugeruje się, że powinna ona trwać podczas całego procesu

eksploracji danych, na wszystkich jego etapach. Głównymi zagadnieniami, jakie powinny zostać na tym etapie zdefiniowane są: zakres analiz lub badań, cel projektu oraz stworzenie planu działania wraz z jego oceną. Należy podkreślić fakt, iż etap potrzeby modelu PDMWD ma charakter organizacyjny, przez co narzuca dalsze warunki przeprowadzenia następnych etapów;

- dane – bazując na sformułowanych wcześniej założeniach, identyfikuje się wymagane źródła danych. W zależności od potrzeb można pracować bezpośrednio na danych źródłowych lub w oparciu o wcześniej przygotowany jednolity wejściowy zbiór danych. Sugeruje się rozważanie rozmiaru wczytywanych danych, gdyż dodatkowa ilość niepotrzebnych danych może stanowić pewne utrudnienie. Z kolei pominięcie istotnych zmiennych może uniemożliwić poprawne przeprowadzenie eksploracji danych. Po uzyskaniu wejściowego zbioru danych proponuje się przegrupowanie struktury danych w celu utworzenia pożądanej struktury, sprawdzenia, oczyszczenia oraz poznania danych. Po pozytywnej ocenie zbioru wejściowego należy stworzyć ostateczny zbiór danych przeznaczony do projektu eksploracji danych, sporządzić raport obejmujący działania związane z identyfikacją źródeł danych, pozyskaniem danych ich wstępną analizą oraz oceną jakości. Dodatkowo na tym etapie proponuje się przeprowadzenie wielowymiarowych analiz na podstawie technologii OLAP. Tego typu analizy pozwalają w pełni poznać przygotowane dane oraz zweryfikować i ocenić ich jakość. Należy szczególnie podkreślić fakt, iż etap ten odgrywa kluczową rolę dla właściwego przeprowadzenia procesu eksploracji danych. Dlatego też, fundamentalną rolę odgrywa tutaj szerokie doświadczenie i duża wiedza z zakresu baz danych, hurtowni danych, itp.;
- modelowanie – etap ten proponuje się rozpocząć od wczytania przygotowanego zbioru danych do aplikacja, z pomocą której będzie realizowana eksploracja danych. W przypadku, gdyby okazało się, że założenia sporządzone na etapie badania potrzeb nie pokrywają się z wymaganiami danej aplikacji, konieczne jest ponowne przygotowanie danych. W przeciwnym razie należy opracować modelowe zbiory danych. Należy dokonać także wyboru, zastosowania oraz parametryzacji techniki bądź metody modelującej. Zaleca się wielokrotne zastosowanie kilku metod i technik do danego zagadnienia eksploracji danych. Utworzony model należy przetestować pod względem poprawności oraz dokonać jego oceny. Sugeruje się, aby wdrożenie modelu poprzedzić ponownym przejrzeniem jego konstrukcji (pod względem poprawności założeń technicznych i biznesowych) oraz przetestowaniem. Etap ten powinien zakończyć się wygenerowaniem raportu wynikowego, sporządzeniem planu monitorowania, utrzymania modelu i dostarczeniu go decydentowi;
- nowa wiedza – na tym etapie uwaga skupiona jest przede wszystkim na przekształceniu wyniku otrzymanego z eksploracji danych w wiedze biznesową. Sugeruje się zapisywanie wyników w takiej formie, aby były one łatwe do przyswojenia przez decydenta. Dobrze w tej roli sprawdzają się metody wizualizacji danych, wśród których należy wymienić m.in. zestawienia tabelaryczne, różnego typu wykresy, grafy oraz mapy. Warto zaznaczyć, iż ocena nowej wiedzy powinna należeć zarówno do decydenta, jak i specjalisty z zakresu eksploracji danych. Wszystkie spostrzeżenia proponuje się zapisać w postaci raportu lub sprawozdania. Warto także dokonać identyfikacji mocnych i słabych stron całego projektu oraz zapisać uwagi dotyczące m.in. takich kwestii, jak na

jakim etapie napotkano znaczne problemy, dlaczego wybrano taką technikę/metodę eksploracji danych, dlaczego w ten sposób zinterpretowano wyniki itp.;

- decyzje – zaprojektowany, oceniony i wdrożony model eksploracji danych należy wykorzystać do opracowania scenariuszy, które powinny umożliwiać rozważenie rezultatów potencjalnych decyzji. Zbudowanie kilku scenariuszy pozwala decydentowi prześledzić różne warianty zdarzeń oraz oszacować następstwa podjęcia różnych decyzji. Ponadto, zachęcają go one do zaznajamiania się z nową wiedzą oraz jej wykorzystaniem. Co warto jeszcze podkreślić to fakt, że stworzone scenariusze mogą okazać się przydatne w przyszłych projektach z zakresu eksploracji danych.

Zaproponowany model PDMWD został tak zaprojektowany, aby umożliwiać decydentowi pozyskanie nieznanej dotąd wiedzy oraz podjęcie na jej podstawie właściwych decyzji. W modelu położono duży nacisk na integrację pozornie odrębnych, aczkolwiek kluczowych, zagadnień związanych z eksploracją danych. Wśród obszarów, jakie uwzględniono w modelu PDMWD, wskazać należy przede wszystkim:

- miejsce i rolę decydenta, a dokładniej wkład jego wiedzy i doświadczenia w złożonym procesie eksploracji danych,
- poznanie potrzeb biznesowych, uwarunkowań, które je implikują,
- rozpoznanie potencjalnych źródeł danych, ich pozyskanie oraz przygotowanie zbioru danych, który zostanie wykorzystany w projekcie eksploracji danych,
- budowę modelu, m.in. poprzez przygotowanie zbioru danych do analizy, wybranie i wykorzystanie stosownych metod, technik eksploracji danych, które pozwolą wydobyć ukrytą wiedzę z przygotowanego zbioru danych,
- interpretację wyników przeprowadzonego procesu eksploracji danych, prowadzącą do ich transformacji w pożądaną wiedzę,
- wykorzystanie pozyskanej wiedzy w procesie podejmowania decyzji gospodarczych,
- utrwalenie i zgromadzenie uzyskanej wiedzy m.in. poprzez sporządzenie i archiwizację potencjalnych scenariuszy.

O ile w branży IT sporo uwagi poświęca się przedsiębiorstwom produkcyjnym, handlowym, o tyle propozycje z zakresu wykorzystania eksploracji danych dla sektora nieruchomości, w tym spółdzielni mieszkaniowych, należą do rzadkości. W związku z powyższym zdecydowano się na zweryfikowanie modelu PDMWD w wybranej spółdzielni mieszkaniowej.

W wyniku przeprowadzenia wywiadu potrzeb informacyjnych wybranej spółdzielni mieszkaniowej stwierdzono, iż jednym z najbardziej niepokojących był fakt wzrostu ilości zgłoszonych drobnych usterek i awarii. W zawiązku z tym, istotnym problemem postawionym przed procesem eksploracji danych, okazało się oszacowanie prawdopodobieństwa wystąpienia awarii dla danej branży oraz oszacowanie kwoty wydatków, związanych z ich usuwaniem jakie mogą pojawić się w przyszłości. Po określeniu zakresu badań przygotowano odpowiedni zbiór danych, który był zgodny z potrzebami projektu. Wybrano określone bazy danych oraz tabele (w tym głównie tabele związane z danymi takimi jak: adresy lokatorów zgłaszających awarię, dane podmiotów usuwających awarię, branża awarii, itd.). Zbiór tych danych został oczyszczony i ujednolicony. W celu uzyskania wstępnych zależności między tymi danymi oraz ich lepszego poznania, opracowano kostki OLAP. Dzięki nim m.in. poznano ilość awarii z danej branży w rozkładzie na spółdzielnię, osiedle, ulicę, blok. Zidentyfikowano także ilość

i charakter napraw, których dokonały konkretne podmioty. Niemniej jednak dopiero po przejściu przez etap modelowanie zauważono, że głównie trzy klasy rozpatrywanych branży tj. elektryczna, hydrauliczna oraz budowlana występują najczęściej. Po dokładnej analizie otrzymanych wyników okazało się, że właśnie te trzy branże wiążą się bezpośrednio z największą ilością interwencji, największymi wygenerowanymi kosztami oraz najdłuższym czasem usuwania awarii dla zidentyfikowanej grupy osiedli. Tego typu uzyskana wiedza była bardzo pożądana przez zarządców spółdzielni mieszkaniowej, gdyż może ona zostać wykorzystana m.in. w planowaniu priorytetów dla przyszłych remontów oraz w pewnym stopniu pozwala określić poziom jakości danego obiektu.

Weryfikacja modelu PDMWD na przykładzie spółdzielni mieszkaniowej pozwoliła na:

- usprawnienie procesów związanych z funkcjonowaniem infrastruktury nieruchomości oraz wydłużeniem cyklu ich życia, poprzez wskazanie obszarów (branża hydrauliczna, elektryczna, ogólnobudowlana itp. z uwzględnieniem bloku lub osiedla, na którym dane prace powinny w pierwszej kolejności zostać wykonane), wymagających przeprowadzenia konserwacji, remontu lub budowy nowych obiektów;
- obniżenie kosztów generowanych przez pojawiające się awarie, gdyż zaplanowanie i przeprowadzenie prac prewencyjnych często wiąże się z niższymi kosztami i niedogodnościami niż usuwanie powstałej awarii wraz z jej skutkami;
- poprawa alokacji posiadanych zasobów (np. planowanie zapasów części zamiennych i materiałów wymaganych przy usuwaniu konkretnych awarii);
- poprawa wizerunku zarządcy (tym samym konkretnej nieruchomości) na rynku nieruchomości. Obniżenie ilości i uciążliwości awarii pozytywnie wpływa na wzrost oceny jakości życia lokatorów danej nieruchomości. Dodatkowo, pośrednio wpływa również na obniżenie ilości pustostanów, czy też rotacji w zawieranych umowach najmu, itp.;
- wzrost zadowolenia najemców z użytkowania danej nieruchomości, który jest wywołany poprawą jej stanu technicznego.

4. Wnioski

Jak wynika z powyższych rozważań eksplorację danych należy traktować wieloaspektowo, kompleksowo i coraz częściej interdyscyplinarnie. Powinna być ona ciągiem wielu etapów mających interakcyjny charakter. Należy podkreślić, że proces ten powinien opierać się na wiedzy, którą posiada nie tylko ekspert z zakresu eksploracji danych, ale również ekspert dziedzinowy. Trzeba zaznaczyć także, że aby przeprowadzenie procesu eksploracji danych w pełni mogło wspierać procesy decyzyjne, konieczne staje się potraktowanie tych pozornie odrębnych zagadnień jako jeden proces. Temu celowi służy model procesowy eksploracji danych PDMWD.

Literatura

1. Bartuś K.: Eksploracja danych. [w:] Strategie i modele gospodarki elektronicznej. Celina M. Olszak, Ewa Ziemba (red.), Wydawnictwo Naukowe PWN, Warszawa, 2007.
2. Berry M.J.A., Linoff G. S.: Nowa Six Sigma. Helion, Gliwice, 2005.
3. Berry M.J.A., Linoff G.S.: Data Mining Techniques For Market Sales and Customer Relationship Management. Willey&Sons, 2004.

4. Demski T.: Jak wdrażać i stosować data mining w praktyce? W: Zastosowanie statystyki i data mining w finansach. StatSoft, Kraków, 2003.
5. Han J., Kamber M.: Data Mining. Concepts and Techniques. Morgan Kaufman Publishers, 2001.
6. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa, 2001.
7. Olszak C. M.: Tworzenie i wykorzystanie systemów Business Intelligence na potrzeby współczesnej organizacji. AE, Katowice, 2007.
8. Shumeli G., Patel N., Bruce P.: Data Mining for Business Intelligence. Wiley, New York, 2007.
9. Smoląg K., Kulej-Dudek E., Pypłacz P. : Zastosowanie Data Mining we współczesnym wspomaganiu zarządzania przedsiębiorstwem. [w:] Kiełtyka L., Nazarko J.: Metody i procesy usprawniania zarządzania przedsiębiorstwem. Wybrane zagadnienia. Wydawnictwo Menedżerskie PTM. Warszawa, 2006.
10. Sokołowski A.: Data Mining – automat czy metoda naukowa? W: Data Mining: poznaj siebie i swoich klientów, StatSoft, Warszawa, 2005.
11. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, USA, 1999.
12. www.gartner.com, 2007.
13. www.kdnuggets.com, 2007.
14. www.sas.com.
15. www.twocrows.com, 2007.

Prof. AE dr hab. inż. Celina M. OLSZAK
Katedra Informatyki Ekonomicznej
Akademia Ekonomiczna im. Karola Adamieckiego
40-287 Katowice, ul. 1 Maja 50
tel./fax.: (0-32) 257 70 00
e-mail: olszak@ekonom.ae.katowice.pl

Dr inż. Kamila BARTUŚ
Katedra Informatyki
Śląska Wyższa Szkoła Zarządzania im. gen. Jerzego Ziętka w Katowicach
40-952 Katowice, ul. Francuska 12
tel./fax.: (0-32) 205 37 45
e-mail: kdymek@swsz.katowice.pl