NAUKA 3/2007 • 83-104

TADEUSZ MORZY

Eksploracja danych

Intensywny rozwój technologii generowania, gromadzenia i przetwarzania danych z jednej strony, z drugiej, upowszechnienie systemów informatycznych, związane ze wzrostem świadomości użytkowników i znaczącym spadkiem cen sprzętu komputerowego, zaowocowały nagromadzeniem olbrzymich wolumenów danych przechowywanych w bazach danych, hurtowniach danych i różnego rodzaju repozytoriach danych. Na ten bezprecedensowy wzrost rozmiaru wolumenów gromadzonych danych w ostatnich latach złożyło się szerokie upowszechnienie narzędzi cyfrowego generowania danych (kodów paskowych, kart płatniczych, aparatów cyfrowych, poczty elektronicznej, sieci RFID, edytorów tekstu, itp.) oraz pojawienie się pojemniejszych i tańszych pamięci masowych. Według raportu UC Berkeley [3], w samym 2002 roku wygenerowano 5 EB (1 EB = 10^{18}) nowych danych. Według przedstawionego raportu, od 2002 roku obserwujemy około 30-procentowy przyrost nowych danych rocznie. Sama poczta elektroniczna generuje, jak się szacuje, około 400 000 TB nowych danych rocznie. Dla porównania, zbiory biblioteki Kongresu USA zawierają około 10 TB danych. Co ciekawe, raport szacuje, że około 90% nowych danych jest gromadzonych na nośnikach magnetycznych, tylko niewielka część nowych danych jest gromadzona na innych nośnikach (film – 7%, papier – 0,01%, nośniki optyczne – 0,002%). Największym „producentem" danych są Stany Zjednoczone – szacuje się, że produkują one około 40% wszystkich danych światowych.

Głównym źródłem danych, co oczywiste, jest bieżąca działalność przedsiębiorstw i instytucji: banków, ubezpieczalni, sieci handlowych, urzędów administracji publicznej i samorządowej itp. Innym dostawcą danych są ośrodki naukowe, które generują olbrzymie ilości danych w każdej niemalże dziedzinie naukowej (fizyka, astronomia, biologia, nauki techniczne itp.). Wiele firm, przedsiębiorstw, instytucji administracji publicznej, ośrodków naukowych, dysponuje bazami i hurtowniami danych o rozmiarach siegających 20-30 TB. Największym repozytorium danych jest w chwili obecnej, oczywiście, sieć Web, zawierająca miliardy stron internetowych (Google indeksuje ponad 8 mld stron, Yahoo indeksuje około 20 mld stron). Archiwum internetowe (*Internet Archive*), utworzone w 1996 r., zgromadziło do chwili obecnej ponad 300 TB danych multimedialnych [13].

Prof. dr hab. Tadeusz Morzy, Instytut Informatyki, Politechnika Poznańska

Nasuwa się naturalne pytanie o celowość przechowywania tak olbrzymich wolumenów danych. Okazuje się, jak wynika z przeprowadzonych badań, że tylko niewielka część zgromadzonych danych jest analizowana w praktyce. Wiele firm i przedsiębiorstw dysponujących zasobami danych, przechowywanych w zakładowych bazach i hurtowniach danych, stanęło przed problemem, w jaki sposób efektywnie i racjonalnie wykorzystać nagromadzoną w tych danych wiedzę dla celów wspomagania swojej działalności biznesowej. Przykładowo, nawet niewielkie sieci supermarketów rejestrują codziennie sprzedaż tysięcy artykułów w kasach fiskalnych. Czy można wykorzystać zgromadzone dane o transakcjach, aby zwiększyć sprzedaż i poprawić rentowność?

Jak już wspomnieliśmy wcześniej, zdecydowana większość danych jest gromadzona na nośnikach magnetycznych w systemach baz i hurtowni danych. Tradycyjny dostęp do tych danych sprowadza się, najczęściej, do realizacji prostych zapytań poprzez predefiniowane aplikacje lub raporty. Sposób, w jaki użytkownik korzysta i realizuje dostęp do bazy danych nazywamy *modelem przetwarzania*. Tradycyjny model przetwarzania danych w bazach danych – „przetwarzanie transakcji w trybie on-line" (ang. *on-line transaction processing*) (OLTP), jest w pełni satysfakcjonujący w przypadku bieżącej obsługi działalności danej firmy, dla dobrze zdefiniowanych procesów (obsługa klienta w banku, rejestracja zamówień, obsługa sprzedaży itp.). Model ten dostarcza efektywnych rozwiązań dla takich problemów, jak: efektywne i bezpieczne przechowywanie danych, transakcyjne odtwarzanie danych po awarii, optymalizacja dostępu do danych, zarządzanie współbieżnością dostępu do danych itd. W znacznie mniejszym stopniu model OLTP wspomaga procesy analizy danych, agregacji danych, wykonywania podsumowań, optymalizacji złożonych zapytań formułowanych *ad hoc* czy wreszcie aplikacji wspomagających podejmowanie decyzji. Prace badawcze i rozwojowe prowadzone nad rozszerzeniem funkcjonalności systemów baz danych doprowadziły do opracowania nowego modelu przetwarzania danych, którego podstawowym celem jest wspomaganie procesów podejmowania decyzji oraz opracowania nowego typu „bazy danych" nazwanego hurtownią danych (ang. *data warehouse*).

Nowy model przetwarzania danych, opracowany dla hurtowni danych, nazwany „przetwarzaniem analitycznym on-line" (ang. *on-line analytical processing*) (OLAP), ma za zadanie wspieranie procesów analizy hurtowni danych dostarczając narzędzi umożliwiających analizę hurtowni w wielu „wymiarach" definiowanych przez użytkowników (czas, miejsce, klasyfikacja produktów itp.). Analiza hurtowni polega na obliczaniu agregatów (podsumowań) dla zadanych „wymiarów" hurtowni. Należy podkreślić, że proces analizy jest całkowicie sterowany przez użytkownika. Mówimy czasami o analizie danych sterowanej zapytaniami (ang. *query-driven exploration*). Typowym przykładem takiej analizy w odniesieniu do hurtowni danych, zawierającej dane dotyczące sprzedaży produktów w supermarkecie jest zapytanie o łączną sprzedaż produktów w kolejnych kwar-

tałach, miesiącach, tygodniach itp., zapytanie o sprzedaż produktów z podziałem na rodzaje produktów (AGD, produkty spożywcze, kosmetyki itp.), czy wreszcie zapytanie o sprzedaż produktów z podziałem na oddziały supermarketu. Odpowiedzi na powyższe zapytania umożliwiają decydentom określenie wąskich gardeł sprzedaży, produktów przynoszących deficyt, zaplanowanie zapasów magazynowych czy porównanie sprzedaży różnych grup produktów w różnych oddziałach supermarketu.

Analiza danych w hurtowni danych, zgodnie z modelem OLAP, jest sterowana całkowicie przez użytkownika. Użytkownik formułuje zapytania i dokonuje analizy danych zawartych w hurtowni. Z tego punktu widzenia, OLAP można interpretować jako rozszerzenie standardu języka dostępu do baz danych SQL o możliwość efektywnego przetwarzania złożonych zapytań zawierających agregaty. Niestety, analiza porównawcza zagregowanych danych, która jest podstawą modelu OLAP, operuje na zbyt szczegółowym poziomie abstrakcji i nie pozwala na formułowanie bardziej ogólnych zapytań: jakie czynniki kształtują popyt na produkty? Czym różnią się klienci supermarketu w Poznaniu i Warszawie? Jakie produkty kupują klienci supermarketu najczęściej wraz z winem? Jakie oddziały supermarketu miały „anormalną" sprzedaż w pierwszym kwartale 2007 roku? Czy można przewidzieć popyt klientów na określone produkty? Czy istnieje korelacja między lokalizacją oddziału supermarketu a asortymentem produktów, których sprzedaż jest wyższa od średniej sprzedaży produktów? Co więcej, „ręczny" charakter analizy OLAP uniemożliwia automatyzację procesu analizy i ogranicza tym samym zakres tej analizy.

Odpowiedzią na potrzebę bardziej zaawansowanej i automatycznej analizy danych, przechowywanych w bazach i hurtowniach danych, jest technologia *eksploracji danych* (ang. *data mining*). Zadaniem metod eksploracji danych, nazywanej również *odkrywaniem wiedzy w bazach danych* (ang. *knowledge discovery in databases, database mining*), jest automatyczne odkrywanie nietrywialnych, dotychczas nieznanych, zależności, związków, podobieństw lub trendów – ogólnie nazywanych *wzorcami* (ang. *patterns*) – w dużych repozytoriach danych. Odkrywane w procesie eksploracji danych wzorce mają, najczęściej postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień, wykresów itp. Automatyczna eksploracja danych otwiera nowe możliwości w zakresie interakcji użytkownika z systemem bazy i (lub) hurtownią danych. Przede wszystkim umożliwia formułowanie zapytań na znacznie wyższym poziomie abstrakcji aniżeli pozwala na to standard SQL. Analiza danych sterowana zapytaniami (OLAP) zakłada, że użytkownik, po pierwsze, posiada pełną wiedzę o przedmiocie analizy i, po drugie, potrafi sterować tym procesem. Eksploracja danych umożliwia analizę danych dla problemów, które, ze względu na swój rozmiar, są trudne do przeprowadzenia przez użytkownika, oraz tych problemów, dla których nie dysponujemy pełną wiedzą o przedmiocie analizy, co uniemożliwia sterowanie procesem analizy danych.

Celem eksploracji danych jest przede wszystkim poznanie i zrozumienie analizowanych procesów i generowanych przez nie danych.

Eksploracja danych jest dziedziną informatyki, która integruje szereg dyscyplin badawczych takich, jak: systemy baz i hurtowni danych, statystyka, sztuczna inteligencja, uczenie maszynowe i odkrywanie wiedzy, obliczenia równoległe, optymalizacja i wizualizacja obliczeń, teoria informacji, systemy reputacyjne. Powyższa lista dyscyplin nie jest pełna. Eksploracja danych wykorzystuje również szeroko techniki i metody opracowane na gruncie systemów wyszukiwania informacji, analizy danych przestrzennych, rozpoznawania obrazów, przetwarzania sygnałów, technologii Web, grafiki komputerowej, bioinformatyki. Jakie dane podlegają eksploracji danych? Początkowo eksploracji poddawano proste typy danych (liczby, łańcuchy znaków, daty), przechowywane w plikach płaskich oraz relacyjnych bazach danych. Wraz z rozwojem narzędzi do generowania i przechowywania danych, z jednej strony, z drugiej, z rozwojem technologii eksploracji danych, eksploracji poddawane są coraz bardziej złożone typy danych: multimedialne (zdjęcia, filmy, muzyka), przestrzenne (mapy), tekstowe i semistrukturalne, przebiegi czasowe, sekwencje danych kategorycznych, grafy, struktury chemiczne (sekwencje DNA), sieci socjalne.

Termin „eksploracja danych" jest często używany jako synonim terminu „odkrywanie wiedzy" w bazach i magazynach danych. W istocie należy rozróżnić dwa pojęcia: odkrywanie wiedzy i eksploracja danych. Zgodnie z definicją [6-9, 24, 26], termin "odkrywanie wiedzy" ma charakter ogólniejszy i odnosi się do całego procesu odkrywania wiedzy, który stanowi zbiór kroków transformujących zbiór danych „surowych" w zbiór wzorców, które mogą być następnie wykorzystane w procesie wspomagania podejmowania decyzji. W procesie odkrywania wiedzy wyróżniamy następujące etapy:

1) Czyszczenie danych (ang. *data cleaning*) – celem etapu jest usunięcie niepełnych, niepoprawnych lub nieistotnych danych ze zbioru eksplorowanych danych;
2) Integracja danych (ang. *data integration*) – celem etapu jest integracja danych z różnych heterogenicznych i rozproszonych źródeł danych w jeden zintegrowany zbiór danych;
3) Selekcja danych (ang. *data selection*) – celem etapu jest selekcja danych istotnych z punktu widzenia procesu analizy danych;
4) Konsolidacja i transformacja danych (ang. *data transformation, data consolidation*) – celem etapu jest transformacja wyselekcjonowanych danych do postaci wymaganej przez metody eksploracji danych;
5) Eksploracja danych (ang. *data mining*) – celem etapu jest odkrywanie potencjalnie użytecznych wzorców ze zbioru wyselekcjonowanych danych.
6) Ocena wzorców (ang. *pattern evaluation*) – celem etapu jest ocena i identyfikacja interesujących wzorców.

7) Wizualizacja wzorców (ang. *knowledge representation*) – celem etapu jest wizualizacja otrzymanych interesujących wzorców w taki sposób, aby umożliwić użytkownikowi interpretację i zrozumienie otrzymanych w wyniku eksploracji wzorców.

W przedstawionym ujęciu, eksploracja danych stanowi tylko jeden z etapów procesu odkrywania wiedzy, którego celem jest odkrywanie wzorców w zbiorze eksplorowanych danych. Celem pozostałych etapów procesu odkrywania wiedzy jest przygotowanie danych, selekcja danych do eksploracji, czyszczenia danych, definiowanie dodatkowej wiedzy przedmiotowej, interpretacja wyników eksploracji i ich wizualizacja. Najczęściej niektóre etapy procesu odkrywania wiedzy są wykonywane łącznie. Przykładowo, etapy czyszczenia danych oraz integracji danych stanowią integralną część budowy hurtowni danych, natomiast etapy selekcji danych, transformacji i konsolidacji danych mogą być zrealizowane poprzez zbiór zapytań. Wzorce odkryte na etapie eksploracji danych są prezentowane użytkownikowi, ale mogą być zapamiętane w bazie danych lub magazynie danych dla dalszej eksploracji.

Oprogramowanie implementujące metody eksploracji danych nazywamy *systemem eksploracji danych* (ang. *data mining system*). W początkowym etapie rozwoju systemów eksploracji danych, algorytmy eksploracji danych były implementowane bądź w postaci niezależnych aplikacji (ang. *stand-alone data mining systems*), bądź bezpośrednio wewnątrz aplikacji użytkownika. Dane do eksploracji były, najczęściej, przechowywane w lokalnych systemach plików użytkownika lub systemach baz danych. W tym drugim przypadku system bazy danych był wyłącznie repozytorium danych, a zaimplementowane algorytmy eksploracji danych nie korzystały z funkcjonalności oferowanej przez system zarządzania bazą danych.

Prace badawcze prowadzone nad rozszerzeniem funkcjonalności systemów baz i hurtowni danych, z jednej strony, oraz zmiana sposobu postrzegania metod eksploracji danych, z drugiej strony, doprowadziły do powstania nowej generacji systemów eksploracji danych. Systemy te, nazywane czasami systemami eksploracji danych drugiej generacji, cechują się silną integracją algorytmów eksploracji danych z podstawową funkcjonalnością systemu zarządzania bazą i (lub) hurtownią danych. Zasadniczymi zaletami takiej integracji są: (1) redukcja kosztów eksploracji danych, (2) wzrost efektywności algorytmów eksploracji danych, (3) wzrost produktywności programistów aplikacji wspomagania podejmowania decyzji oraz (4) wzrost bezpieczeństwa danych i ochrony danych przed nieuprawnionym dostępem. Redukcja kosztów eksploracji danych wynika z redukcji kosztów przygotowania danych do eksploracji, to jest, czyszczenia danych, integracji danych i konsolidacji danych – etapy te stanowią integralną część procesu budowy bazy/hurtowni danych. Wzrost efektywności algorytmów eksploracji danych wynika z możliwości wykorzystania wewnętrznych mechanizmów systemu zarządzania bazą danych do wspierania i poprawy efektywności procesu eksploracji danych takich,

jak: indeksy, perspektywy materializowane, buforowanie danych, wstępne sprowadzanie danych do bufora, optymalizacja wykonywania zapytań do bazy danych. Wzrost produktywności programistów aplikacji wspomagania podejmowania decyzji wynika z dwóch przesłanek. Po pierwsze, z faktu, że eksploracja danych stanowi integralną część systemu zarządzania bazą danych i może być postrzegana jako usługa systemowa. Po drugie, z faktu, że wzorce otrzymane w wyniku eksploracji danych są przechowywane w bazie danych i mogą stanowić argument innych usług systemowych, na przykład, aplikacji raportujących. Wreszcie, wzrost bezpieczeństwa i ochrony danych wynika również z możliwości wykorzystania wewnętrznych mechanizmów systemu zarządzania bazą danych, takich jak transakcyjne odtwarzanie bazy danych po awarii czy autoryzacja dostępu do danych. Zagadnienie ochrony danych i odkrytych wzorców przed nieuprawnionym dostępem nabrało w ostatnim czasie szczególnego znaczenia w kontekście problemu ochrony prywatności. Wyniki niektórych metod eksploracji danych mogą naruszać prywatność osób fizycznych. Dotyczy to eksploracji baz danych DNA, danych medycznych, historii korzystania z kart kredytowych, historii dostępu do stron WWW itd. W tym kontekście bardzo intensywnie są prowadzone, w ostatnim czasie, prace nad algorytmami eksploracji danych, zapewniającymi ochronę prywatności (ang. *privacy-preserving data mining*) [2, 25]. Systemowe wsparcie dla tej klasy algorytmów eksploracji danych stanowi dodatkowy argument przemawiający za integracją metod eksploracji danych i systemów zarządzania bazami/hurtowniami danych.

Systemy eksploracji danych drugiej generacji oferują użytkownikom programowy interfejs, którym jest deklaratywny „język zapytań eksploracyjnych" (ang. *data mining query* language), nazywany również językiem eksploracji danych [9, 12, 19]. Język zapytań eksploracyjnych umożliwia użytkownikom specyfikację: zadania eksploracji danych, zbioru eksplorowanych danych, odkrywanych wzorców oraz szczegółowych ograniczeń dla odkrywanych wzorców. Proces eksploracji danych ze swej natury jest procesem interakcyjnym i iteracyjnym. Korzystając z języka eksploracji danych, użytkownik definiuje zadanie eksploracji danych, a następnie, po otrzymaniu i przeanalizowaniu uzyskanych wzorców, ma możliwość zredefiniowania parametrów tego zadania przez zawężenie/rozszerzenie zbioru eksplorowanych danych lub zawężenie/rozszerzenie zbioru poszukiwanych wzorców. Przykładowo, załóżmy, że użytkownik definiuje następujące zadanie odkrywania asocjacji: „znajdź wszystkie zbiory produktów występujących wspólnie w co najmniej 20% koszyków klientów supermarketu w okresie ostatnich 5 lat". Jeżeli w wyniku eksploracji danych otrzymany zbiór wzorców jest, na przykład, zbyt liczny, użytkownik może zawęzić zbiór eksplorowanych danych oraz poszukiwanych wzorców: „znajdź wszystkie zbiory produktów występujących wspólnie w co najmniej 25% koszyków klientów supermarketu w okresie ostatnich 3 lat". Co więcej, wynik eksploracji użytkownik może zapamiętać w bazie/hurtowni danych dla późniejszej analizy porów-

nawczej – „jak zmienił się koszyk wspólnych produktów kupowanych w tym roku w porównaniu z rokiem ubiegłym".

Różnorodność i wielość metod eksploracji danych, wywodzących się często z różnych dyscyplin badawczych, utrudnia potencjalnym użytkownikom identyfikację metod, które są najodpowiedniejsze z punktu widzenia ich potrzeb w zakresie analizy danych. Metody eksploracji danych można sklasyfikować, ogólnie, ze względu na cel eksploracji, typy eksplorowanych danych oraz typy wzorców odkrywanych w procesie eksploracji danych. Najpopularniejszą klasyfikacją metod eksploracji danych jest klasyfikacja tych metod ze względu na cel eksploracji. Z tego punktu widzenia metody eksploracji danych można podzielić, bardzo ogólnie, na następujące klasy [9, 19, 24]:

- Odkrywanie asocjacji – najszersza klasa metod, obejmująca, najogólniej, metody odkrywania interesujących zależności lub korelacji, nazywanych ogólnie asocjacjami, między danymi w dużych zbiorach danych. Wynikiem działania metod odkrywania asocjacji są zbiory reguł asocjacyjnych opisujących znalezione zależności i (lub) korelacje.
- Klasyfikacja i predykcja – obejmuje metody odkrywania modeli (tak zwanych klasyfikatorów) lub funkcji opisujących zależności pomiędzy zadaną klasyfikacją obiektów a ich charakterystyką. Odkryte modele klasyfikacji są następnie wykorzystywane do klasyfikacji nowych obiektów.
- Grupowanie (analiza skupień, klastrowanie) – obejmuje metody znajdowania skończonych zbiorów klas obiektów posiadających podobne cechy. W przeciwieństwie do metod klasyfikacji i predykcji, klasyfikacja obiektów (podział na klasy) nie jest znana a priori, lecz jest celem metod grupowania. Metody te grupują obiekty w klasy w taki sposób, aby maksymalizować podobieństwo wewnątrzklasowe obiektów i minimalizować podobieństwo pomiędzy klasami obiektów.
- Analiza sekwencji i przebiegów czasowych – obejmuje metody analizy sekwencji danych kategorycznych (np. sekwencji biologicznych), sekwencji zbiorów danych kategorycznych oraz przebiegów czasowych. Metody analizy sekwencji danych mają na celu znajdowanie częstych podsekwencji (tzw. wzorców sekwencji, częstych epizodów), klasyfikację i grupowanie sekwencji, natomiast metody analizy przebiegów czasowych mają na celu głównie znajdowanie trendów, podobieństw, anomalii oraz cykli w przebiegach czasowych.
- Odkrywanie charakterystyk – obejmuje metody znajdowania zwięzłych opisów lub podsumowań ogólnych własności klas obiektów. Znajdowane opisy mogą mieć postać reguł charakteryzujących lub reguł dyskryminacyjnych. W tym drugim przypadku opisują różnice między ogólnymi własnościami tak zwanej klasy docelowej (klasy analizowanej) a własnościami tak zwanej klasy (zbioru klas) kontrastującej (klasy porównywanej).

- Eksploracja tekstu i danych semistrukturalnych – obejmuje metody reprezentacji i analizy danych tekstowych oraz danych semistrukturalnych (XML) w celu ich grupowania, klasyfikacji oraz wspierania procesu wyszukiwania.
- Eksploracja WWW – obejmuje metody analizy korzystania z sieci Web w celu znajdowania typowych wzorców zachowań użytkowników sieci, metody analizy powiązań stron w sieci Web, w celu określenia ważności i koncentratywności stron, a tym samym, poprawy efektywności procesu wyszukiwania stron, metody grupowania i klasyfikacji stron WWW, w oparciu o ich zawartość i schemat zewnętrzny, wreszcie, metody analizy ukrytych sieci socjalnych, „stron lustrzanych" i wewnętrznych „środowisk" (ang. *communities*) oraz ich ewolucję. W ostatnim czasie bardzo intensywnie rozwijaną grupą metod eksploracji WWW stanowią metody analizy reklam internetowych (ich efektywności, rozliczania i propagacji).
- Eksploracja grafów i sieci socjalnych – obejmuje metody analizy struktur grafowych oraz sieci socjalnych. Struktury te są aktualnie szeroko wykorzystywane do modelowania złożonych obiektów takich, jak: obwody elektroniczne, związki chemiczne, struktury białkowe, sieci biologiczne, sieci socjalne, procedury obiegu dokumentów, powiązania między uczestnikami gier i aukcji internetowych, dokumenty XML itp. Popularną grupę metod eksploracji struktur grafowych stanowią algorytmy odkrywania częstych podstruktur (podgrafów) w bazie danych struktur grafowych. Ważną grupę metod stanowią algorytmy klasyfikacji struktur grafowych, umożliwiające znajdowanie zależności między pewną charakterystyką struktury grafowej a jej budową (np. analiza i klasyfikacja sekwencji DNA). Inną, bardzo intensywnie rozwijaną w ostatnim czasie grupą algorytmów są algorytmy analizy sieci socjalnych, wspomagające: procesy wykrywania oszustów uczestniczących w aukcjach internetowych, wykrywanie przestępstw w kryminalistyce, analizę dużych sieci elektrycznych i telekomunikacyjnych itp.
- Eksploracja danych multimedialnych i przestrzennych – obejmuje metody analizy i eksploracji multimedialnych oraz przestrzennych baz danych, przechowujących obrazy, mapy, dźwięki, wideo itp. Zasadniczym celem metod eksploracji danych multimedialnych jest wspomaganie procesów wyszukiwania danych. Metody eksploracji danych multimedialnych, służące do grupowania i klasyfikacji danych, są najczęściej silnie powiązane z mechanizmami systemu zarządzania bazą danych (indeksowanie i buforowanie danych.
- Wykrywanie punktów osobliwych – obejmuje metody wykrywania (znajdowania) obiektów osobliwych, które odbiegają od ogólnego modelu danych (klasyfikacja i predykcja) lub modeli klas (analiza skupień). Często metody wykrywania punktów osobliwych stanowią integralną część innych metod eksploracji danych, np. metod grupowania.

Przejdziemy obecnie do omówienia kilku podstawowych i najpopularniejszych metod eksploracji danych.

Metody eksploracji danych

Odkrywanie asocjacji

Odkrywanie asocjacji jest jedną z najciekawszych i najbardziej popularnych technik eksploracji danych. Celem procesu odkrywania asocjacji jest znalezienie interesujących zależności lub korelacji, nazywanych ogólnie asocjacjami między danymi w dużych zbiorach danych [1, 9-11, 19-23].

Wynikiem procesu odkrywania asocjacji jest zbiór reguł asocjacyjnych opisujących znalezione zależności między danymi. Początkowo problem odkrywania asocjacji był rozważany w kontekście tak zwanej „analizy koszyka zakupów" (ang. *market basket analysis*). Klasyczny problem analizy koszyka zakupów polega na analizie danych pochodzących z kas fiskalnych i opisujących zakupy realizowane przez klientów w supermarkecie. Celem tej analizy jest znalezienie naturalnych wzorców zachowań konsumenckich klientów przez analizę produktów, które są przez klientów supermarketu kupowane najczęściej wspólnie (tj. określenie grup produktów, które klienci najczęściej umieszczają w swoich koszykach – stąd nazwa problemu). Znalezione wzorce zachowań klientów mogą być następnie wykorzystane do opracowania akcji promocyjnych, organizacji półek w supermarkecie, opracowania koncepcji katalogu oferowanych produktów itp.

Rozważmy, dla ilustracji, uproszczony przykład bazy danych supermarketu, przedstawionej na rycinie 1, zawierającej informacje o transakcjach zrealizowanych przez klientów supermarketu. Każda transakcja składa się ze zbioru produktów. Z każdą transakcją klienta jest związany unikalny identyfikator transakcji oraz data realizacji transakcji.

trans_id	produkty	data
1	chleb, mleko	02/22/9
2	chleb, pieluszki, piwo	02/22/98
3	mleko, pieluszki, piwo, ser	02/23/98
4	chleb, mleko, piwo, pieluszki	02/24/98
5	chleb, mleko, ser, pieluszki	02/24/98

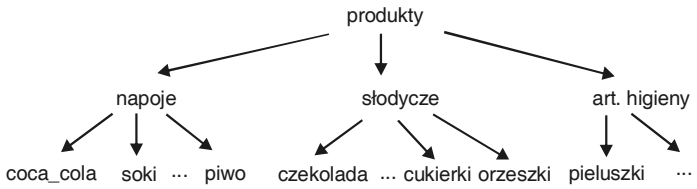
Ryc. 1. Przykładowa baza danych supermarketu

W bazie danych, przedstawionej na rycinie 1, można znaleźć regułę asocjacyjną postaci: „60% klientów, którzy kupują pieluszki, kupuje również piwo". Wiedza, jaką niesie powyższa reguła, może być wykorzystana dwojako: do reorganizacji produktów na pół-

kach sklepowych oraz do organizacji akcji promocyjnej. Umieszczenie piwa i pieluszek obok siebie na półkach może zwiększyć sprzedaż obu produktów, odwołując się do naturalnych preferencji konsumenckich klientów. Innym przykładem wykorzystania powyższej reguły asocjacyjnej może być akcja promocyjna piwa, prowadzona pod hasłem 10% obniżki ceny piwa. Naturalną konsekwencją akcji powinno być zwiększenie liczby klientów kupujących piwo, a tym samym zrekompensowanie straty wynikającej z oferowanej obniżki ceny piwa. Jednakże dodatkowym źródłem dochodu może być również zwiększony zysk z tytułu dodatkowej sprzedaży pieluszek kupowanych najczęściej łącznie z piwem.

W literaturze poświęconej eksploracji danych można znaleźć wiele rodzajów reguł asocjacyjnych [9, 19, 22-24]. Reguły te można sklasyfikować według szeregu kryteriów: typu przetwarzanych danych, wymiarowości przetwarzanych danych lub stopnia abstrakcji przetwarzanych danych. Wyróżniamy binarne i ilościowe reguły asocjacyjne, jedno- i wielowymiarowe reguły asocjacyjne oraz jedno- i wielopoziomowe reguły asocjacyjne. Przykładem ilościowej reguły asocjacyjnej jest reguła postaci: „5% osób uprawnionych do głosowania w USA, w wieku między 30. a 40. rokiem życia, z wykształceniem wyższym, głosuje na demokratów". Regułę asocjacyjną nazywamy ilościową regułą asocjacyjną, jeżeli dane występujące w regule są danymi ciągłymi i (lub) kategorycznymi.

Czasami dane występujące w bazie danych tworzą pewną hierarchię poziomów abstrakcji, którą nazywamy taksonomią (patrz ryc. 2)



Ryc. 2. Taksonomia produktów

Przykładowo, produkty w supermarkecie można poklasyfikować według kategorii produktu: produkt 'pieluszki_Pampers' należy do kategorii 'pieluszki', która, z kolei, należy do kategorii 'art.higieny'; produkt 'piwo_żywiec' należy do kategorii 'piwo', która, z kolei należy do kategorii 'napoje'; wreszcie, produkt 'czekolada' należy do kategorii 'słodycze'. Reguły, które opisują asocjacje, występujące pomiędzy danymi reprezentującymi różne poziomy abstrakcji, nazywamy *wielopoziomowymi regułami asocjacyjnymi*. Przykładem wielopoziomowej reguły asocjacyjnej jest reguła postaci: „10% klientów, którzy kupują artykuły higieniczne i czekoladę, kupuje również jakiś napój". Zauważmy, że dane występujące w powyższej regule reprezentują różne poziomy abstrakcji: dana „art. higieny" reprezentuje wyższy poziom abstrakcji aniżeli dana opisująca produkt

„czekolada". W ostatnim czasie zdefiniowano szereg nowych typów reguł asocjacyjnych: cykliczne, nieoczekiwane, rozmyte, negatywne, dysocjacyjne itd. Przegląd i pełną klasyfikację reguł asocjacyjnych można znaleźć w pracach [9, 24].

Odkrywanie asocjacji znalazło zastosowanie w wielu dziedzinach takich jak: handel (analiza koszyka zakupów, organizacja akcji promocyjnych), analiza dokumentów (wspomaganie wyszukiwania dokumentów), analiza sieci telekomunikacyjnych (wykrywanie sytuacji alarmowych, wielowymiarowa analiza danych, analiza opłacalności usług telekomunikacyjnych), wykrywanie intruzów w sieciach komputerowych, bioinformatyka (analiza sekwencji DNA, sekwencji białkowych). Odkrywanie asocjacji jest wykorzystywane również szeroko w innych metodach eksploracji danych: klasyfikacji, predykcji czy też grupowaniu.

Klasyfikacja

Klasyfikacja jest metodą analizy danych, której celem jest przypisanie obiektu (danej) do jednej z predefiniowanych klas w oparciu o zbiór wartości atrybutów opisujących dany obiekt. Mówiąc inaczej, celem klasyfikacji jest predykcja wartości pewnego wybranego atrybutu kategorycznego obiektu, nazywanego atrybutem decyzyjnym, w oparciu o zbiór wartości atrybutów opisujących dany obiekt, tak zwanych deskryptorów. Wartości atrybutu decyzyjnego dzielą zbiór obiektów na predefiniowane klasy, składające się z obiektów o tej samej wartości atrybutu decyzyjnego. Klasyfikacja znajduje bardzo szerokie zastosowanie w szeregu dziedzinach: bankowości, medycynie, przetwarzaniu tekstów, biologii itp. Przykłady zastosowania klasyfikacji obejmują: (1) klasyfikację pacjentów do zabiegu operacyjnego w oparciu o atrybuty opisujące stan pacjenta, (2) procedurę przyznawania kredytu bankowego w oparciu o charakterystykę klienta, (3) wykrywanie i klasyfikację wiadomości e-mailowych typu „spam", (4) automatyczną klasyfikację kierowców na powodujących i niepowodujących wypadków drogowych [4, 9, 14, 21, 26].

Najpopularniejszym podejściem do klasyfikacji obiektów jest konstrukcja modeli, nazywanych klasyfikatorami, które każdemu obiektowi przydzielają wartość atrybutu decyzyjnego w oparciu o wartości deskryptorów. Punktem wyjścia do konstrukcji klasyfikatora jest zbiór obiektów historycznych, będący listą wartości deskryptorów i wybranego atrybutu decyzyjnego. Klasyfikacja danych, z wykorzystaniem klasyfikatora, jest procesem dwuetapowym. W kroku pierwszym jest konstruowany klasyfikator opisujący predefiniowany zbiór klas obiektów. W kroku drugim, opracowany klasyfikator jest stosowany do predykcji wartości atrybutu decyzyjnego (klasy) obiektów, dla których wartość atrybutu decyzyjnego, tj. przydział do klasy, nie jest znany. Model klasyfikacyjny (klasyfikator) jest również budowany dwuetapowo. Zbiór obiektów historycznych, nazywany również zbiorem przykładów, obserwacji lub próbek, dzielimy na dwa zbiory: zbiór treningowy i zbiór testowy. W kroku pierwszym budowy klasyfikatora (faza

uczenia lub treningu), w oparciu o zbiór treningowy danych jest konstruowany klasyfikator. W kroku drugim, nazywanym fazą testowania, w oparciu o zbiór testowy danych jest szacowana dokładność (jakość) klasyfikatora.

Testowanie jakości klasyfikatora polega na obliczeniu współczynnika dokładności, który jest obliczany jako procent przykładów testowych poprawnie zaklasyfikowanych przez klasyfikator. Klasyfikator ma najczęściej postać drzew decyzyjnych, reguł klasyfikacyjnych lub formuł logicznych. Przykłady reguł klasyfikacyjnych: „kierowcy prowadzący czerwone pojazdy o pojemności 650 cm^2 powodują wypadki drogowe", „kierowcy, którzy posiadają prawo jazdy ponad 3 lata lub jeżdżą niebieskimi samochodami nie powodują wypadków drogowych".

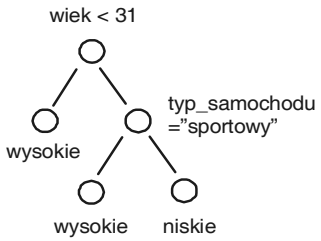
Dla ilustracji rozważmy bardzo prosty przykład bazy danych ubezpieczalni, przedstawiony na rycinie 3. Baza danych zawiera informacje o klientach ubezpieczalni i zawiera cztery atrybuty: „klient_id", „wiek", „typ_samochodu" oraz „Ryzyko". Atrybut „Ryzyko" jest atrybutem decyzyjnym i dzieli wszystkich klientów na dwie klasy: ryzyko „wysokie" – klient o wysokim stopniu ryzyka (powodujący wypadki); ryzyko „niskie" – klient o niskim stopniu ryzyka (niepowodujący wypadków). Atrybut „klient_id" jest kluczem przedstawionej tabeli (unikalnym identyfikatorem każdego wiersza tabeli), pozostałe atrybuty: „wiek" oraz „typ_samochodu", są deskryptorami.

klient_id	wiek	typ_samochodu	ryzyko
1	20	kombi	wysokie
2	18	sportowy	wysokie
3	40	sportowy	wysokie
4	50	sedan	niskie
5	35	minivan	niskie
6	30	kombi	wysokie
7	32	sedan	niskie
8	40	kombi	niskie
9	27	sportowy	wysokie
10	34	sedan	niskie
11	66	sedan	wysokie
12	44	sportowy	wysokie

Ryc. 3. Przykładowa baza danych ubezpieczalni

Rekordy bazy danych, przedstawionej na rycinie 3, zostały podzielone na dwa zbiory: zbiór danych treningowych (rekordy o numerach 1-8) oraz zbiór danych testowych (rekordy o numerach 9-12). Rycina 4 przedstawia klasyfikator zbudowany w oparciu o zbiór danych treningowych.

Klasyfikator ten ma postać drzewa decyzyjnego. Równoważny zapis klasyfikatora, w postaci zbioru reguł klasyfikacyjnych, przedstawiono na rycinie 5.



Ryc. 4. Przykład klasyfikatora

if wiek < 31 then Ryzyko ="wysokie"
if wiek ≥ 31 and typ_samochodu = „sportowy" then Ryzyko ="wysokie"
if wiek ≥ 31 and typ_samochodu ≠ „sportowy" then Ryzyko ="wysokie"

Ryc. 5. Zbiór reguł klasyfikacyjnych

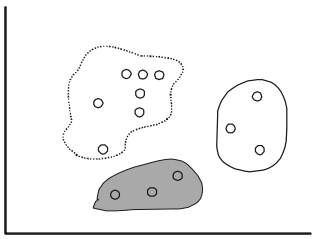
W kolejnym kroku, w oparciu o zbiór danych testowy (rekordy 9-12), jest szacowana dokładność (jakość) klasyfikatora. Zauważmy, że jakość klasyfikatora wynosi 75%, gdyż klasyfikator poprawnie klasyfikuje 3 z 4 przypadków – rekordy 9, 10 i 12. Klasyfikator błędnie klasyfikuje rekord 11.

Istnieje wiele metod klasyfikacji, różniących się dokładnością, odpornością na szum i błędy w danych treningowych, skalowalnością, wreszcie, interpretowalnością [9, 14, 21, 24]. Do najpopularniejszych metod klasyfikacji można zaliczyć, wspomnianą wcześniej klasyfikację poprzez indukcję drzew decyzyjnych, naiwny klasyfikator Bayesa, sieci bayesowskie, sieci neuronowe, SVM (ang. *Support Vector Machine*), kNN. Klasyfikacji mogą podlegać obiekty opisane zarówno danymi ciągłymi, jak i danymi kategorycznymi, sekwencje danych kategorycznych i danych ciągłych, sekwencje zbiorów, dane tekstowe i dane semistrukturalne, struktury grafowe, utwory muzyczne, filmy itp.

Grupowanie

Inną, bardzo popularną metodą eksploracji danych jest grupowanie. Pod pojęciem grupowania rozumiemy proces grupowania obiektów, rzeczywistych bądź abstrakcyjnych, o podobnych cechach, w klasy nazywane klastrami lub skupieniami [7-9, 15-17]. Ideę grupowania ilustruje rycina 6.

Istnieje wiele różnych definicji pojęcia klastra: (1) zbiór obiektów, które są „podobne", (2) zbiór obiektów, takich że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza aniżeli odległość między dowolnym obiektem należącym do klastra i dowolnym obiektem nienależącym do tego klastra, czy też (3) spójny obszar przestrzeni wielowymiarowej charakteryzujący się dużą gęstością występowania obiektów.



Ryc. 6. Grupowanie danych

Grupowanie, podobnie jak klasyfikacja, znajduje bardzo szerokie zastosowanie w wielu dziedzinach: bankowości, medycynie, przetwarzaniu tekstów, biologii itp. Przykłady zastosowania grupowania:

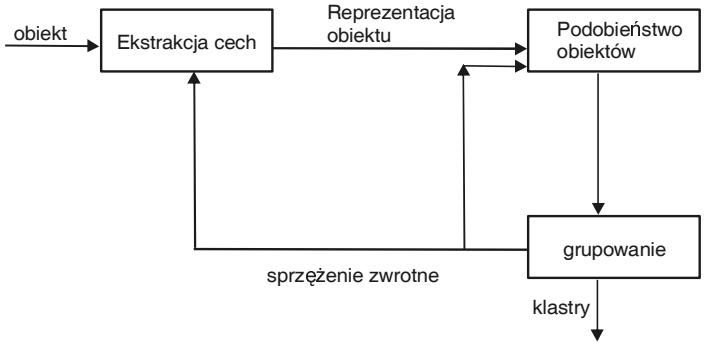
- grupowanie zbiorów dokumentów tekstowych, grupowanie przychodzących wiadomości e-mailowych – zbiór dokumentów (zbiór wiadomości) interpretujemy jako zbiór punktów w przestrzeni wielowymiarowej, w której pojedynczy wymiar odpowiada jednemu słowu z określonego słownika. Współrzędne dokumentu w przestrzeni wielowymiarowej są zdefiniowane względną częstością występowania słów ze słownika. Klastry dokumentów odpowiadają grupom dokumentów dotyczących podobnej tematyki,
- grupowanie zbiorów sekwencji dostępu do stron WWW – pojedyncza sekwencja opisuje sekwencję dostępów do stron WWW danego serwera, realizowaną w ramach jednej sesji przez użytkownika. Klastry sekwencji odpowiadają grupom użytkowników danego serwera, którzy realizują dostęp do tego serwera w podobny sposób,
- grupowanie klientów (sklepy Amazon, Merlin) – każdy klient jest opisany zbiorem zakupionych produktów. Klastry produktów odpowiadają grupom klientów, którzy charakteryzują się podobnymi upodobaniami.
- grupowanie pacjentów według objawów – każda choroba charakteryzuje się szeregiem objawów. Grupowanie pacjentów o podobnych objawach można wykorzystać do identyfikacji różnych typów chorób. Grupowanie wykorzystano, m.in. do identyfikacji różnych typów depresji odpowiadających różnym objawom.

Proces grupowania obiektów składa się z kilku kroków, które często stanowią osobne metody eksploracji danych, znajdujące zastosowanie w innych metodach przetwarzania danych (patrz rycina 7).

Podstawowymi krokami procesu grupowania są:

1) Wybór reprezentacji obiektów – procedura ekstrakcji/selekcji najbardziej istotnych cech z punktu widzenia procesu grupowania obiektów;
2) Wybór miary podobieństwa pomiędzy obiektami – bardzo silnie zależy od dziedziny zastosowań i grupowanych typów danych;

3) Grupowanie obiektów (klastry) – wybór określonego algorytmu grupowania obiektów;
4) Znajdowanie charakterystyki klastrów – znajdowanie zwięzłego oraz czytelnego dla użytkownika opisu klastrów.



Ryc. 7. Elementy składowe procesu grupowania

Krok pierwszy, wybór reprezentacji obiektów, ma na celu selekcję tych cech opisujących obiekty, które są istotne z punktu widzenia procesu grupowania obiektów. Przykładowo, z punktu widzenia procesu grupowania klientów sklepu internetowego, istotnymi cechami są: wysokość zapłaconych dotychczas rachunków, lokalizacja klienta, charakterystyka zakupionych produktów (np. w koszyku zakupów klienta przeważają płyty DVD oraz przewodniki turystyczne) czy też adres URL klienta; natomiast drugorzędne znaczenie mają atrybuty opisujące samego klienta (wiek, płeć czy też kolor oczu). Krok drugi, wybór miary podobieństwa między obiektami, ma na celu wybór najlepszej miary określającej podobieństwo grupowanych obiektów. Miara ta silnie zależy od dziedziny zastosowań i grupowanych typów danych. Jeżeli grupowane obiekty można przetransformować do punktów w przestrzeni euklidesowej, to za miarę podobieństwa dwóch obiektów można przyjąć odległość między tymi obiektami (odległość euklidesową, odległość blokową, odległość Minkowskiego). W przypadku gdy obiekty nie poddają się transformacji do przestrzeni euklidesowej, proces grupowania wymaga zdefiniowania innych miar odległości (podobieństwa) między obiektami. Dotyczy to niestandardowych obiektów typu: sekwencje dostępów do stron WWW, sekwencje DNA, sekwencje zbiorów, zbiory atrybutów kategorycznych, dokumenty tekstowe, XML, struktury grafowe, wykresy EKG, wideo, dźwięki itp.

Istnieje wiele różnych metod i algorytmów grupowania. Metody i algorytmy grupowania można sklasyfikować ze względu na: (1) metodę przeszukiwania przestrzeni stanów wszystkich możliwych partycji zbioru grupowanych obiektów – metody deterministyczne i probabilistyczne, (2) metodę konstrukcji klastrów – metody hierarchiczne i po-

działowe, (3) charakter znajdowanych klastrów – metody generujące klastry rozłączne i metody generujące klastry przecinające się, (4) metodę wykorzystywania cech obiektów w procesie grupowania – metody monoatrybutowe i poliatrybutowe, (5) typ grupowanych danych – metody grupowania danych liczbowych, kategorycznych, sekwencji, struktur grafowych itp., (6) metodę odświeżania klastrów – metody przyrostowe (klastry są tworzone w sposób przyrostowy, wraz ze wzrostem liczby obiektów do grupowania) lub proces grupowania jest realizowany od początku; problem ten dotyczy dużych zbiorów obiektów i, dodatkowo, występują ograniczenia na czas grupowania oraz dostępną pamięć.

Odkrywanie charakterystyk

Celem metod odkrywania charakterystyk jest znajdowanie zwięzłych opisów (charakterystyk) podanego zbioru danych. Przykładem takiej charakterystyki jest opis pacjenta chorującego na anginę: „pacjenci chorujący na anginę cechują się temperaturą ciała większą niż 37,5 C, bólem gardła oraz ogólnym osłabieniem organizmu". Opisy zbiorów danych są znajdowane w dwojaki sposób przez: (1) odkrywanie charakterystyki zbioru danych, (2) analizę dyskryminacyjną zbioru danych lub (3) połączenie obu technik. Odkrywanie charaktrystyki zbioru danych, nazywanego często zbiorem celowym (ang. *target class*), polega na podsumowaniu danych należących do podanego zbioru. Zbiór celowy danych jest najczęściej uzyskiwany poprzez realizację zapytania do bazy lub hurtowni danych. Przykładowo, w celu znalezienia spółek giełdowych, których ceny akcji rosną co miesiąc o 10%, należy wykonać zapytanie do bazy danych zawierającej informacje o wartościach akcji spółek giełdowych. Następnie, uzyskany zbiór danych poddajemy dalszej analizie w celu znalezienia zwięzłej charakterystyki tego zbioru. Istnieje szereg efektywnych metod znajdowania podsumowań i charakterystyk zbiorów celowych. Większość tych metod została opracowana dla potrzeb analitycznego przetwarzania danych (OLAP) w hurtowniach danych. Wyniki odkrywania charakterystyk zbiorów danych mają najczęściej postać wykresów graficznych lub reguł charakterystycznych (tzw. *characteristic rules*) [7, 9].

Analiza dyskryminacyjna zbioru danych polega na porównania podstawowych cech zbioru celowego z cechami zbioru (lub zbiorów) porównawczych, nazywanych zbiorami kontrastującymi (ang. *contrasting classes*). Zbiór celowy danych, jak i zbiory kontrastujące, są uzyskiwane przez realizację zapytania do bazy lub hurtowni danych. Przykładowo, użytkownik systemu eksploracji danych może być zainteresowany porównaniem spółek giełdowych, których ceny akcji rosną co miesiąc o 10%, ze spółkami giełdowymi, których ceny akcji maleją w tym samym czasie o 10%. Oba zbiory danych uzykujemy przez realizację odpowiednich zapytań do bazy danych, zawierającej informacje o wartościach akcji spółek giełdowych. Następnie oba uzyskane zbiory danych poddajemy dal-

szej analizie, stosując metody podobne do tych stosowanych w odkrywaniu charakterystyk. Podobnie, wyniki analizy dyskryminacyjnej zbiorów danych mają najczęściej postać wykresów lub reguł dyskryminacyjnych (tzw. *discriminant rules*) [7-9].

Metody odkrywania charakterystyk są stosowane bądź jako niezależne metody eksploracji danych, bądź stanowią element innej metody eksploracji danych, np. metody grupowania w celu znalezienia zwięzłych opisów uzyskanych klastrów. Przykładowo, w ramach procesu grupowania klientów sklepu internetowego, gdzie każdy klient jest opisany zbiorem zakupionych produktów, możemy być zainteresowani znalezieniem zwięzłej charakterystyki klientów, którzy każdorazowo realizują zakupy powyżej 100 zł. Inne ciekawe przykłady zastosowań można znaleźć w pracy [9].

Eksploracja sieci Web

Najdynamiczniej rozwijającym się w ostatnim czasie obszarem badawczym, w zakresie eksploracji danych, są metody i algorytmy eksploracji sieci Web [5, 9, 18, 24]. Eksploracja sieci Web to odkrywanie interesującej, potencjalnie użytecznej, dotychczas nieznanej wiedzy (reguł, wzorców, zależności) ukrytej w zawartości sieci Web i sposobie korzystania z niej. Sieć Web jest bardzo specyficznym repozytorium danych. Sieć Web przypomina bazę danych, ale przechowywane dane (tj. strony WWW) są nieustrukturalizowane, a złożoność danych jest znacznie większa aniżeli złożoność tradycyjnych dokumentów tekstowych. Ponadto, pojedyncza dana stanowi połączenie dokumentu tekstowego, czasami z elementami multimedialnymi, oraz zbioru hiperlinków (połączeń) do innych danych. Dane opisujące korzystanie z sieci Web, przechowywane w tzw. logach serwerów WWW, mają bardzo duże rozmiary i bardzo dynamiczny przyrost. Dzienny przyrost danych przechowywanych w logach serwerów firmy Google jest porównywalny z największymi konwencjonalnymi hurtowniami danych. Sieć Web jest bardzo dynamicznym środowiskiem, jednakże tylko bardzo niewielka część informacji zawartej w sieci Web jest istotna dla pojedynczego użytkownika. To wszystko sprawia, że eksploracja sieci Web stanowi olbrzymie wyzwanie badawcze, a wyniki badań znajdują natychmiastowe zastosowanie w praktyce.

Jak już wspomnieliśmy, sieć Web można rozpatrywać jako specyficzne repozytorium danych nieustrukturalizowanych. Stąd w odniesieniu do sieci Web i jej zawartości informacyjnej znajdują zastosowanie wszystkie wcześniej wspomniane metody eksploracji danych. Niemniej dla potrzeb eksploracji sieci Web zaproponowano szereg nowych, całkowicie oryginalnych i specyficznych metod i algorytmów eksploracji danych.

Najogólniej, metody eksploracji sieci Web można podzielić na trzy grupy [5, 9, 18]: (1) eksploracja zawartości sieci Web (ang. *Web content mining*), (2) eksploracja połączeń sieci Web (ang. *Web linkage mining*) oraz (3) eksploracja korzystania z sieci Web (ang. *Web usage mining*). Pierwsza grupa metod i algorytmów ma na celu wspieranie

i poprawę efektywności procedur wyszukiwania zawartości sieci Web. Do tej grupy metod należą metody wyszukiwania stron WWW, opracowane dla języków języki zapytań do sieci Web (WebSQL, WebOQL, WebML, WebLog, W3QL), algorytmy grupowania i klasyfikacji stron WWW, oraz bardzo specyficzne algorytmy eksploracji for dyskusyjnych i ocen klientów w celu określenia ich preferencji. Część zadań związanych z eksploracją zawartości sieci to tradycyjne zadania eksploracji danych, np. grupowanie i klasyfikacja stron WWW – wykorzystuje się w tym celu algorytmy grupowania i klasyfikacji dokumentów XML. Część zadań to zadania specyficzne. Do tej grupy należą algorytmy eksploracji for dyskusyjnych i blogów. Druga grupa metod i algorytmów, tj. algorytmy eksploracji połączeń sieci Web, analizują sieć połączeń (hiperlinków), reprezentującą strukturę sieci Web. Analiza struktury sieci Web pozwala na określanie rankingu ważności stron WWW, co stanowi podstawowy element wyszukiwarek internetowych (Google, Yahoo, ASK), znajdowanie lustrzanych serwerów WWW, znajdowanie grup użytkowników o podobnych zainteresowaniach itp. Najbardziej znane algorytmy eksploracji sieci Web, pozwalające na określanie rankingu wyszukiwanych stron, to PageRank oraz algorytm HITS. Algorytmy eksploracji sieci Web swoje źródło mają w pracach nad analizą sieci socjalnych (ang. *social network analysis*). Aktualnie, poza wyszukiwarkami internetowymi, algorytmy eksploracji sieci połączeń są wykorzystywane z powodzeniem w systemach rekomendacyjnych (Netflix, Amazon), reputacyjnych, systemach obsługi aukcji internetowych – w celu określania reputacji uczestników aukcji, w kryminalistyce – w celu określania powiązań osób podejrzanych. Celem trzeciej grupy metod i algorytmów, tj. algorytmów eksploracji korzystania z sieci Web, jest analiza danych opisujących korzystanie z zasobów sieci Web w celu znajdowania ogólnych wzorców zachowań użytkowników sieci, w szczególności odkrywania wzorców dostępu do stron WWW. Dane opisujące dostęp użytkowników do zasobów sieci są pamiętane w logo serwerów WWW. Wynikiem eksploracji logo serwerów WWW są ogólne wzorce dostępu do stron WWW. Najpopularniejszą grupą algorytmów, stosowaną do eksploracji logo WWW, są algorytmy odkrywania wzorców sekwencji. Odkryta wiedza pozwala na: (1) budowę adaptatywnych serwerów WWW, co pozwala na personalizację usług serwerów WWW i dostosowanie ich do potrzeb i wymagań grup użytkowników (systemy rekomendacyjne, handel elektroniczny np. Amazon), (2) optymalizację struktury serwera i poprawę nawigacji po zasobach serwera WWW (zastosowanie – reorganizacja serwisu Yahoo), (3) znajdowanie potencjalnie najlepszych miejsc reklamowych w sieci, wreszcie (4) poprawę efektywności procedur wstępnego ściągania stron na żądanie (ang. *prefatching*), co istotnie wpływa na efektywność działania systemów baz danych, hurtowni danych oraz serwerów WWW wykorzystujących te procedury.

Reasumując, metody eksploracji znalazły bardzo liczne zastosowania w sieci Web, począwszy od wspomagania działania wyszukiwarek sieciowych (Google, Yahoo, Ask),

grupowania i klasyfikacji stron WWW, przez handel elektroniczny (systemy rekomendacyjne, odkrywanie asocjacji), reklamy internetowe (*Google Adsense*), wykrywanie oszustw i analizę reputacji kupujących i sprzedających na aukcjach internetowych, projektowanie serwerów WWW (personalizacja usług, adaptatywne serwery WWW), analizę sieci socjalnych, a skończywszy na optymalizacji działania systemów baz danych, hurtowni danych i serwerów WWW.

Problemy odkrywania wiedzy

Eksploracja danych to proces automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców, podobieństw lub trendów w dużych repozytoriach danych (bazach danych, hurtowniach danych itp.). Innymi słowy, celem eksploracji danych jest analiza danych i procesów w celu lepszego ich rozumienia.

Algorytmy eksploracji danych znajdują zależności, wzorce, podobieństwa lub trendy w danych historycznych. Pamiętać jednakże należy, że przeszłość nie jest tożsama z przyszłością, a co za tym idzie, wnioskowanie, na podstawie danych historycznych, o trendach, które potencjalnie mogą wystąpić w przyszłości, może być obarczone poważnym błędem. Innym, znanym problemem związanym z eksploracją danych jest problem wywodzenia wiedzy racjonalnej, uogólnionej z danych empirycznych. Dane historyczne nie muszą opisywać, i najczęściej nie opisują, wszystkich możliwych przypadków, które mogą wystąpić w rzeczywistości. Załóżmy, że dokonaliśmy eksploracji danych pochodzących z kas fiskalnych supermarketu zlokalizowanego w regionie zamieszkałym głównie przez osoby o wysokich dochodach. Przeniesienie i zastosowanie uzyskanych reguł asocjacyjnych w odniesieniu do supermarketu zlokalizowanego w regionie zamieszkałym przez osoby o niskich dochodach nie będzie miało najmniejszego sensu, gdyż będzie prowadziło do błędnych decyzji biznesowych.

Odkrywane reguły asocjacyjne nie implikują „przyczynowości" zdarzeń opisanych tymi regułami. Wspomniana już wcześniej, i często cytowana w literaturze, reguła asocjacyjna odkryta w repozytorium Wal-Martu „60% klientów, którzy kupują pieluszki, kupuje również piwo", nie implikuje, że istnieje jakaś zależność o charakterze przyczynowo-skutkowym pomiędzy kupowaniem pieluszek a kupowaniem piwa. Ta reguła stwierdza jedynie, że wiele osób, które dokonywały zakupów w sieci Wal-Mart w analizowanym okresie, które kupiły pieluszki, kupiły również piwo, a współwystępowanie obu produktów w koszykach klientów było statystycznie znaczące. Tylko tyle i aż tyle. Nie oznacza to, że klienci innych sieci handlowych, w innym okresie czasu, będą mieli podobne preferencje zakupowe.

Poza wspomnianymi problemami natury „jakościowej", systemy eksploracji danych napotykają na problemy natury „ilościowej". W dużych bazach i hurtowniach danych mo-

gą zostać odkryte tysiące reguł. Przykładowo, dla bazy danych o rozmiarze rzędu 100 tys. rekordów, liczba wygenerowanych reguł asocjacyjnych jest o rząd wielkości większa.

Użytkownik ma najczęściej olbrzymie problemy z analizą tak dużych wolumenów danych. Zdecydowana większość otrzymanych reguł ma często niewielką przydatność praktyczną – większość reguł jest znana użytkownikom, będącymi ekspertami w danej dziedzinie przedmiotowej. Na pytanie: „w jaki sposób system eksploracji danych, odkrywając reguły asocjacyjne, może określić, które ze znalezionych reguł są interesujące dla użytkownika?", odpowiedź brzmi: „nie istnieją obiektywne kryteria pozwalające określić wartość (przydatność) znalezionych reguł – tylko użytkownik potrafi ocenić, na ile znaleziona reguła jest interesująca!"

Ostatni problem natury społecznej, związany z eksploracją danych, o którym już wspominaliśmy wcześniej, to problem ochrony prywatności. W roku 1980 OECD (*Organization for Economic Co-operation and Development*) ustanowiła zbiór zaleceń dotyczących ochrony prywatności i jakości przetwarzanych danych (tzw. *fair information practices*). Zalecenia te odnoszą się do problemów związanych ze zbieraniem danych, ich przetwarzaniem, bezpieczeństwem oraz ich dostępnością. Wraz ze wzrostem wolumenów danych przechowywanych i przetwarzanych w wersji elektronicznej oraz dostępnością narzędzi do eksploracji tych danych szczególnego znaczenia nabrało zagadnienie bezpieczeństwa, ochrony i poufności danych. Eksploracja danych może naruszać prywatność osób fizycznych i mieć istotne reperkusje społeczne dla tych osób. Dotyczy to, jak już wspominaliśmy, eksploracji baz danych sekwencji DNA, danych medycznych, historii kart kredytowych, dostępu do stron WWW itp.

Można bowiem sobie łatwo wyobrazić, że w wyniku eksploracji bazy danych sekwencji DNA znaleziono regułę określającą współwystępowanie określonej sekwencji DNA i zawału serca. Nie można zagwarantować, że nie pojawi się pokusa u pracodawców, aby wymagać w momencie zatrudnienia informacji o DNA pracownika i w konsekwencji nie zatrudniać osób, którym grozi zawał serca.

W tym kontekście, w ostatnim czasie, obserwujemy intensywne prace nad rozwojem i propagowaniem systemów eksploracji danych zapewniających ochronę prywatności. Podstawowym zaleceniem jest, aby eksploracji poddawać tylko takie dane, które uniemożliwiają, na ich podstawie, identyfikację osób fizycznych, których one dotyczą.

Piśmiennictwo

[1] Agrawal, R., Srikant, R., *Mining Sequential Patterns*, Proc. 11th International Conference on Data Engineering, 1995, 3-14.

[2] Agrawal, R., Srikant, R., *Privacy-Preserving data Mining*, Proc. ACM SIGMOD Conference on Management of Data, 2000, 439-450.

[3] Berkeley Report, *How Much Information? 2003*, http://www.sims.berkeley. edu/research/projects/how-much-info-2003, 2003.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and regression trees*, Wadsworth, Belmont, 1984.
[5] Chakrabarti, S., *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Pub., 2003.
[6] Chen, M.S., Han, J., Yu, P.S., *Data mining: an overview from a database perspective*, IEEE Trans. Knowledge and Data Engineering, 8, 1996, 866-883.
[7] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., *Advances in knowledge discovery and data mining*, MIT Press, 1996.
[8] Hand, D., Mannila, H., Smyth, P., *Eksploracja danych*, WNT, 2005.
[9] Han, J., Kamber, M., Data mining: concepts and techniques, Morgan Kaufmann Pub., 2006.
[10] Han, J., Pei, J., Mortazavi-Asl, B. et al. *FreeSpan: frequent pattern-projected sequential pattern mining*, Proc. 6th International Conference on Knowledge Discovery and Data Mining (KDD'00), 2000, 355-359.
[11] Han, J., Pei, J., Yin, Y., *Mining frequent patterns without candidate generation*, Proc. 2000 ACM-SIGMOD International Conference on Management of Data, 2000, 1-12.
[12] Imielinski, T., Mannila, H, *A database perspective on knowledge discovery*, Communications of ACM, 39, 1996, 58-64.
[13] Internet Archiwum, http://www.archive.org, 2007.
[14] James, M., *Classification algorithms*, John Wiley, New York, 1985.
[15] Jain, A.K., Dubes, R.C., *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
[16] Jain, A.K., Murty, M.N., Flynn, P.J., *Data clustering: a survey*, ACM Computing Surveys, 1999, 264-323.
[17] Kaufman, L., Rousseeuw, P.J., *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, 1990.
[18] Liu, B., *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 2007.
[19] Morzy, T., *Odkrywanie asocjacji: algorytmy i struktury danych*, OWN, 2004.
[20] Pei, J., Han J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu M-C., *PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth*, Proc. 17th International Conference on Data Engineering (ICDE'01), 2001, 215-224.
[21] Quinlan, J.R., *Induction of decision trees*, Machine Learning, vol. 1, No. 1, 1986, 81-106.
[22] Srikant, R., Agrawal, R., Mining generalized association rules, Proc. 21th International Conference on Very Large Data Bases (VLDB'95), 1995, 407-419.
[23] Srikant, R., Agrawal, R., *Mining quantitative association rules in large relational tables*, Proc. 1996 ACM-SIGMOD International Conference on Management of Data, 1996, 1-12.
[24] Tan, P-N., Steinbach, M., Kumar, V., *Introduction to Data mining*, Pearson Education, 2006.
[25] Verykios, V.S., Bertino, E., Fovino, I.N. et al. *State-of-the-art in privacy preserving data mining*, SIGMOD Record, 33, 1, 2004, 50-57.
[26] Witten, I.H., Frank, E., *Data mining: practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann Pub., 2000.

Data mining

Recent advances in data capture, data transmission and data storage technologies have resulted in a growing gap between more powerful database systems and users' ability to understand and effectively analyze the information collected. Many companies and organizations gather gigabytes or terabytes of business transactions, scientific data, web logs, satellite pictures, text

reports, which are simply too large and too complex to support a decision making process. Traditional database and data warehouse querying models are not sufficient to extract trends, similarities and correlations hidden in very large databases. The value of the existing databases and data warehouses can be significantly enhanced with help of data mining. Data mining is a new research area which aims at nontrivial extraction of implicit, previously unknown and potentially useful information from large databases and data warehouses. Data mining, also referred to as database mining or knowledge discovery in databases, can help answer business questions that were too time consuming to resolve with traditional data processing techniques. The process of mining the data can be perceived as a new way of querying – with questions such as "which clients are likely to respond to our next promotional mailing, and why?". The aim of this paper is to present an overall picture of the data mining field as well as presents briefly few data mining methods. Finally, we summarize the concepts presented in the paper and discuss some problems related with data mining technology.

Key words: data mining, data analysis, evolution of information technology, association analysis, classification, clustering, Web mining