V Konferencja PLOUG
Zakopane
Październik 1999

Eksploracja danych: problemy i rozwiązania

Tadeusz Morzy
morzy@put.poznan.pl
Instytut Informatyki
Politechnika Poznańska

Streszczenie

Artykuł zawiera krótką genezę i opis aktualnego stanu rozwoju ważnej i bardzo intensywnie rozwijanej w ostatnim czasie dziedziny eksploracji danych. Artykuł zawiera krótki przegląd metod eksploracji danych, związków eksploracji danych z magazynami i systemami baz danych, oraz prezentuje możliwe dziedziny zastosowań technik eksploracji danych.

1. Wstęp

Eksploracja danych (ang. data mining), nazywana często potocznie *odkrywaniem wiedzy w bazach danych* (ang. knowledge discovery in databases), jest jedną z najdynamiczniej i najintensywniej rozwijanych dziedzin informatyki w ostatnim czasie. Integruje wiele dyscyplin takich jak: statystyka, systemy baz danych, sztuczna inteligencja, optymalizacja, obliczenia równoległe. Olbrzymie zainteresowanie eksploracją danych wynika z faktu, że szereg przedsiębiorstw, instytucji administracji publicznej czy wreszcie ośrodków naukowych nagromadziło w ostatnim czasie bardzo wiele danych przechowywanych w zakładowych bazach danych i stanęło przed problemem, w jaki sposób efektywnie i racjonalnie wykorzystać nagromadzoną w tych bazach wiedzę dla celów wspomagania swojej działalności biznesowej.

Tradycyjny sposób korzystania z baz danych sprowadza się, najczęściej, do realizacji zapytań poprzez aplikacje lub raporty. Sposób w jaki użytkownik korzysta z bazy danych (w jaki realizuje do niej dostęp) nazywamy *modelem przetwarzania*. Tradycyjny model przetwarzania – *„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.). Niestety, ten klasyczny model przetwarzania danych nie wspomaga procesów analizy danych oraz aplikacji wspomagających podejmowanie decyzji.

Szereg przedsiębiorstw dysponuje olbrzymimi bazami danych. Dysponując danymi w bazie danych opisującymi działalność dużego supermarketu w dłuższym przedziale czasu (sprzedaż produktów, zamówienia, stan rezerw) możemy postawić szereg pytań:
W jaki sposób wykorzystać przechowywane dane do usprawnienia funkcjonowania firmy? Jakie czynniki kształtują taki a nie inny 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 1999 r? Czy można przewidzieć przyszłe zachowania klientów? Dane przechowywane w bazie danych zawierają w sobie potencjalnie olbrzymią wiedzę o otaczającym świecie.

Niestety, ciągle jeszcze niedostatecznie umiemy dokonać analizy tych danych i uzyskać dostęp do zawartej w nich wiedzy. Istniejące interfejsy pomiędzy użytkownikami baz danych a bazami danych ciągle jeszcze nie wspomagają w dostatecznym stopniu nawigowania, podsumowywania, analizy czy modelowania bardzo dużych baz danych. Opracowanie i dostarczenie użytkownikom nowych interfejsów wspomagających wymienione wyżej funkcje jest zadaniem i celem badań prowadzonych w zakresie systemów magazynów i eksploracji danych.

2. OLAP – weryfikacja hipotez

Komercyjnie dostępne systemy transakcyjne (systemy zarządzania bazami danych SZBD) dostarczają efektywnych rozwiązań dla takich problemów jak: efektywne i bezpieczne przechowywanie danych, transakcyjne odtwarzanie danych, dostępność danych, optymalizacja dostępu do danych, zarządzanie współbieżnością. W znacznie mniejszym stopniu systemy te wspomagają operacje agregacji danych, wykonywania pewnych podsumowań czy też optymalizacji złożonych zapytań formułowanych ad hoc. W ostatnim czasie 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 relacyjnej bazy danych nazwanego *magazynem danych* (ang. data warehouse).

Nowy model przetwarzania danych, nazwany „*przetwarzaniem analitycznym on-line*" (ang. On Line Analytical Processing OLAP), ma za zadanie wspieranie procesów analizy magazynów danych dostarczając narzędzi umożliwiających analizę magazynu w wielu „wymiarach" definiowanych przez użytkowników (czas, miejsce, klasyfikacja produktów, itp.). Analiza magazynu polega na obliczaniu agregatów dla zadanych „wymiarów" magazynu. 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 jest zapytanie o sprzedaż produktów w supermarkecie w kolejnych kwartałach, 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, itp., oraz podjęcie odpowiednich działań poprawiających sytuację.

3. Eksploracja danych – odkrywanie hipotez

Analiza danych w magazynie danych, zgodnie z modelem OLAP, jest sterowana całkowicie przez analityka. Analityk formułuje zapytania i dokonuje analizy danych zawartych w magazynie. Z tego punktu widzenia, OLAP można interpretować jako rozszerzenie standardu SQL o możliwości efektywnego przetwarzania złożonych zapytań zawierających agregaty.

W przeciwieństwie do technologii OLAP, technologia eksploracji danych umożliwia automatyczną analizę i eksploracje danych. **Problem eksploracji danych polega na efektywnym znajdowaniu nieznanych dotychczas zależności i związków pomiędzy danymi**. Automatyczna eksploracja danych otwiera nowe możliwości w zakresie interakcji użytkownika z systemem bazy danych (lub magazynem danych). Przede wszystkim umożliwia formułowanie zapytań na znacznie wyższym poziomie abstrakcji aniżeli pozwala na to standard SQL. Analiza danych sterowana zapytaniami, charakterystyczna dla technologii 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 człowieka oraz tych problemów, dla których nie dysponujemy pełną wiedzą – tę wiedzę chcemy wydobyć z danych.

To drugie zagadnienie wiąże się bezpośrednio z *problemem formułowania zapytań*: w jaki sposób uzyskać dostęp do danych w przypadku kiedy nie potrafimy sformułować zapytania w terminach języka dostępu do bazy danych? Jest to typowa sytuacja w systemach wspomagania podejmowania decyzji. Przykładowo, w jaki sposób zidentyfikować rekordy w bazie danych firmy telekomunikacyjnej, które odpowiadają „fałszywym" połączeniom? Podobnie, w przypadku kart kredytowych interesuje nas wykrycie kradzieży tych kart i ich niestandardowe wykorzystanie. W przypadku analizy danych naukowych uzyskanych z dużej liczby eksperymentów interesuje nas wykrycie ciekawych przypadków. Oczywiście, można analizować rekord po rekordzie w bazie danych rozpatrując oddzielnie każdy przypadek; podejście takie jest jednak mało realistyczne w przypadku giga i tera bajtowych baz danych. Z drugiej strony, bardzo trudno sformułować zapytanie w języku SQL, lub nawet zdefiniować procedurę składowaną, które umożliwiłoby przeprowadzenie takiej analizy.

4. Metody eksploracji danych

Jak już wspomnieliśmy na wstępie termin eksploracja danych jest często używany jako synonim procesu odkrywania wiedzy w bazach danych. W literaturze czasami jednak rozróżnia się te dwa pojęcia. Zgodnie z definicją [2] termin odkrywanie wiedzy odnosi się do całego procesu, natomiast eksploracja danych stanowi tylko jeden z etapów tego procesu odnoszący się do generowania reguł.

Pozostałe etapy procesu odnoszą się do przygotowania danych, wyboru danych do eksploracji, czyszczenia danych, definiowania dodatkowej wiedzy przedmiotowej, interpretacji wyników eksploracji i ich wizualizacji.

Metody eksploracji danych można podzielić, bardzo ogólnie, na 6 zasadniczych klas.

- **Odkrywanie asocjacji**
  Najszersza klasa metod obejmująca, najogólniej, odkrywanie różnego rodzaju nieznanych zależności w bazie danych. Metody te obejmują głównie odkrywanie asocjacji pomiędzy obiektami. Generalnie, odkrywane zależności posiadają pewne miary statystyczne określające ich wsparcie i ufność.
- **Klastrowanie**
  Celem tych metod jest znajdowanie skończonego zbioru klas obiektów (klastrów) w bazie danych posiadających podobne cechy. Liczba klastrów jest nieznana, stąd, proces klastrowania przebiega, najczęściej, w dwóch cyklach: cykl zewnętrzny przebiega po liczbie możliwych klastrów, cykl wewnętrzny próbuje znaleźć optymalny podział obiektów pomiędzy klastry.
- **Odkrywanie wzorców sekwencji**
  Odkrywanie czasowych wzorców zachowań, np. znajdowanie sekwencji notowań giełdowych, zachowań klientów ubezpieczalni, klientów supermarketów.
- **Odkrywanie klasyfikacji**
  Celem tych metod jest znajdowanie zależności pomiędzy klasyfikacją obiektów (klasyfikacja naturalna bądź wprowadzona przez eksperta) a ich charakterystyką. Zastosowanie: charakterystyka pacjentów, klientów kart kredytowych, pożyczkobiorców.
- **Odkrywanie podobieństw w przebiegach czasowych**
  Znajdowanie podobieństw w przebiegach czasowych opisujących określone procesy.
- **Wykrywanie zmian i odchyleń**
  Znajdowanie różnic pomiędzy aktualnymi a oczekiwanymi wartościami danych: znajdowanie anomalnych zachowań klientów ubezpieczalni, klientów kart kredytowych, klientów firm telekomunikacyjnych.

5. Odkrywanie asocjacji

Dane:
- $I=\{i1, i2, ..., in\}$ – zbiór obiektów
- Transakcja T: zbiór obiektów takich, że $T \subseteq I$
- Baza danych D: zbiór transakcji
- Transakcja T zawiera X, gdzie $X \subseteq I$, jeżeli $X \subseteq T$
- Reguła asocjacyjna: implikacja postaci $X \Rightarrow Y$, gdzie $X, Y \subseteq I$
- Reguła $X \Rightarrow Y$ posiada zaufanie c% w bazie danych D jeżeli c% transakcji, należących do D i zawierających X zawiera również Y.
- Reguła $X \Rightarrow Y$ posiada wsparcie s w bazie danych D jeżeli s% transakcji należących do D zawiera $X \cup Y$.

Sformułowanie problemu:

> Znajdź wszystkie reguły asocjacyjne w D, których wsparcie s > minsup, i zaufanie c > minconf, gdzie minsup i minconf wartości zadane przez użytkownika.

Przykład:

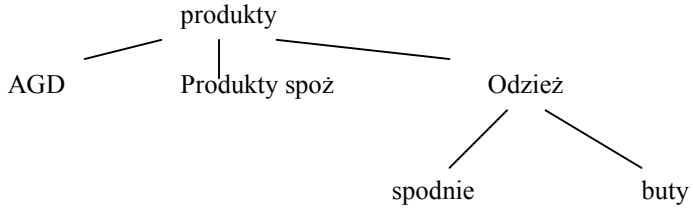
Transakcja	Obiekty
100	A, B, C
200	A, C
300	A, D
400	B, E, F

Dla minsup=50% i minconf=50% następujące reguły asocjacyjne są prawdziwe:

A ⇒ C, reguła posiada 50% wsparcie i 66.6% zaufanie
C ⇒ A, reguła posiada 50% wsparcie i 100% zaufanie

Zastosowania: analiza koszyka zakupów, bezpośredni marketing,

Obiekty mogą tworzyć hierarchie:



Reguły asocjacyjne uwzględniające hierarchie obiektów nazywamy **uogólnionymi regułami asocjacyjnymi:**

spodnie ⇒ AGD

Reguły asocjacyjne odkrywane z danych zarówno numerycznych jak i symbolicznych nazywamy **ilościowymi regułami asocjacyjnymi.**

ID	Wiek	Zarobek	Status małż.	Ilość samoch.
100	44	1000	Żonaty	2
101	55	2000	Żonaty	3

Przykład ilościowej reguły asocjacyjnej:

10% żonatych mężczyzn w wieku 40 – 60 posiada co najmniej dwa samochody w rodzinie

6. Odkrywanie wzorców sekwencji

Dane:

- I={i1, i2, ..., in} – zbiór obiektów
- Transakcja T: zbiór obiektów takich, że T⊆ I
- Sekwencja: lista transakcji pojedynczego klienta
- Baza danych sekwencji D: zbiór sekwencji

Sformułowanie problemu:

Dana jest baza danych sekwencji D. Znajdź wszystkie maksymalne podsekwencje w D, których wsparcie s > minsup, gdzie minsup wartość zadana przez użytkownika. Znalezione podsekwencje nazywamy **wzorcami sekwencji**.

Przykład:

ID klienta	Data	Obiekty
100	3 maja	A
100	15 maja	A
200	4 maja	A
200	16 maja	B
200	2 września	C, E
300	4 kwietnia	A, E
400	3 maja	A
400	3 czerwca	C, D, E
400	2 września	A
500	5 września	A

Dla minsup=40% następujące wzorce sekwencji są spełnione:

A poprzedza A, wzorzec posiada wsparcie 40%
A poprzedza C i E, wzorzec posiada wsparcie 40%

Zastosowania: analiza dostępów do stron w Web, analiza koszyka zakupów, bezpośredni marketing, medycyna, ubezpieczenia, telekomunikacja

Użytkownik może być zainteresowany wzorcami sekwencji spełniającymi zadane ograniczenia. Ograniczenia mogą mieć charakter ograniczeń nałożonych na dane (ang. item constraints), na przykład - znajdź wszystkie produkty, które poprzedzają najczęściej kupno pralki, lub ograniczeń czasowych (ang. time constraints), na przykład – jakie produkty klienci kupują najczęściej w przeciągu miesiąca. Podobnie jak w przypadku reguł asocjacyjnych, produkty mogą tworzyć mogą hierarchie. Wzorce sekwencji uwzględniające hierarchie obiektów nazywamy **uogólnionymi wzorcami sekwencji:**

Kupno telewizora poprzedza kupno produktu AGD

7. Odkrywanie klasyfikacji

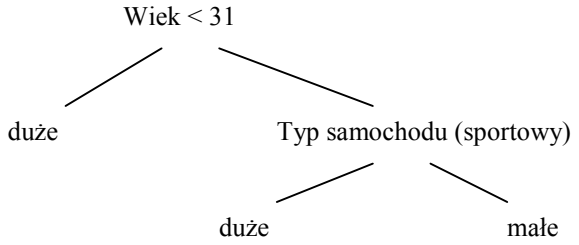
Sformułowanie problemu:

Dana jest baza danych rekordów. Każdy rekord posiada etykietę klasy, do której należy. Znajdź model każdej klasy, tj. opis rekordów dla każdej z klas.

Przykład:

Wiek	Typ samochodu	Ryzyko wypadku
20	Kombi	duże
18	Sportowy	duże
40	Sportowy	duże
50	Kombi	małe
35	Minivan	małe
30	Kombi	duże

Wynik klasyfikacji można przedstawić w postaci tzw. drzewa decyzyjnego. Dla powyższego zbioru rekordów drzewo decyzyjne ma następującą postać:



Odkrywanie klasyfikacji znajduje zastosowanie w takich dziedzinach jak: klasyfikacja pacjentów, analiza wiarygodności kredytobiorców, lokalizacja sklepów, marketing celowy, itp. Problem klasyfikacji jest problemem znanym i analizowanym od wielu lat, szczególnie, w dziedzinie sztucznej inteligencji i uczenia maszynowego. Na gruncie tych dziedzin opracowano szereg metod odkrywania klasyfikacji takich jak: drzewa decyzyjne, sieci neuronowe, algorytmy kombinatoryczne (szczególnie - algorytmy genetyczne), algorytmy statystyczne. Niestety, algorytmy te charakteryzują się słabą skalowalnością. W przeciwieństwie do nich, algorytmy odkrywania klasyfikacji opracowane w ramach eksploracji danych nie mają tej wady; są skalowalne i nie wprowadzają ograniczeń na liczbę rekordów, atrybutów czy też klas.

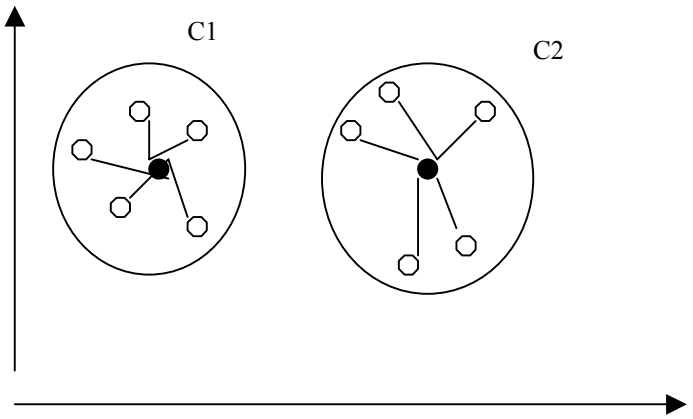
Zastosowania: klasyfikacja pacjentów, weryfikacja kredytobiorców, lokalizacja sklepów, marketing bezpośredni.

8. Klastrowanie

Problem klastrowania wiąże się z zagadnieniem klasyfikacji. W przypadku klasyfikacji, obiekty (rekordy) są przypisane do klas. W przypadku klastrowania poszukujemy klas obiektów o zbliżonych cechach. Sformułowanie problemu:

Dana jest baza danych rekordów. Przypisz rekordy do klastrów.

Czym jest klaster rekordów? Klaster jest podzbiorem „podobnych" rekordów. Można przytoczyć również inne definicje klastrów. Klaster jest podzbiorem rekordów takich, że odległość pomiędzy dwoma dowolnymi rekordami w klastrze jest mniejsza aniżeli odległość pomiędzy dowolnym rekordem w klastrze a rekordem z innego klastra. Inna definicja: klastrem nazywamy spójny obszar w wielowymiarowej przestrzeni o dużej gęstości rekordów.



Problem klastrowania danych, czasami nazywany taksonomią danych, jest analizowany już od wielu lat w ramach sztucznej inteligencji i uczenia maszynowego. Istnieje wiele algorytmów klastrowania danych. Różnice pomiędzy algorytmami wynikają z charakteru danych: dane ciągłe, numeryczne czy też symboliczne. Wyróżnia się dwa zasadnicze typy algorytmów klastrowania danych: algorytmy podziału i algorytmy hierarchiczne. Podobnie jak w przypadku algorytmów odkrywania klasyfikacji, zasadniczym problemem w przypadku stosowania algorytmów opracowanych na gruncie uczenia maszynowego jest ich niewielka skalowalność w zakresie rozmiaru przestrzeni.

Zastosowania: segmentacja rynku klientów(telekomunikacja, ubezpieczenia), segmentacja obrazów, biologia, medycyna.

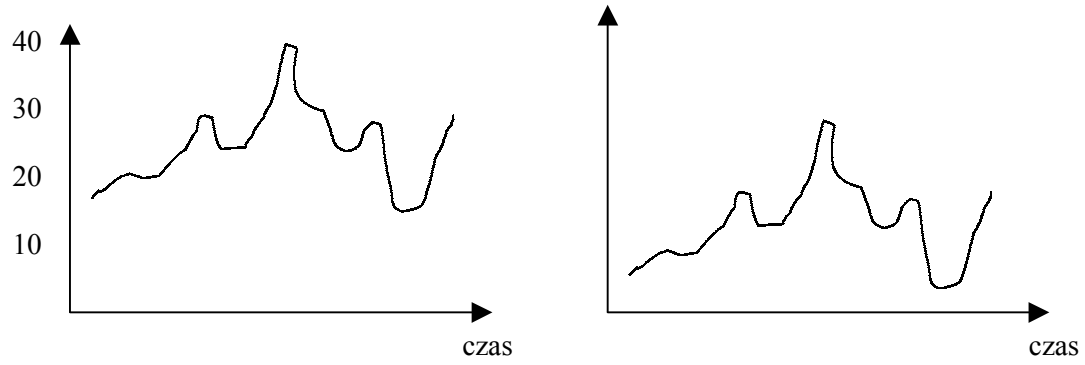
9. Odkrywanie podobieństw w przebiegach czasowych

Dane:

- Baza danych przebiegów czasowych

Sformułowanie problemu:

> Dana jest baza danych przebiegów czasowych. Pogrupuj wszystkie „podobne" przebiegi czasowe.



Zyski z funduszu ALFA Zyski z funduszu BETA

Przedstawione przebiegi czasowe różnią się wartościami i są wzajemnie przesunięte w czasie, mają jednak podobną charakterystykę. Do znajdowania podobieństw czasowych wykorzystuje się transformaty: Fouriera i falkową.

Zastosowania: znajdowanie klientów o podobnej konsumpcji energii elektrycznej, identyfikacja firm na giełdzie o podobnej dynamice wzrostu cen akcji, identyfikacja surowców o podobnej charakterystyce sprzedaży.

10. Wykrywanie zmian i odchyleń

Użytkownicy są zainteresowani w takim samym stopniu odkrywaniem zależności i prawidłowości w bazach danych jak i odkrywaniem odchyleń, anomalii od normalnych zachowań czy też wyjątków. Odchyleniem nazywamy różnicę pomiędzy aktualną a oczekiwaną wartością.

Nazwa	**Styczeń**	**Luty**	**Marzec**	***Kwiecień***
AIX	20	24	30	*36*
NT	10	13	16	*20*
HP	5	6	8	*9*
SUN	30	40	47	*33*

Nazwa	**Styczeń**	**Luty**	**Marzec**	***Kwiecień***
AIX	20	24	30	*28*
NT	10	13	16	*14*
HP	5	6	8	*6*
SUN	30	40	47	*33*

Wykrywanie zmian i odchyleń jest stosowane obecnie głównie do analizy dużych wolumenów wielowymiarowych danych. Analizując dane staramy się zrozumieć trendy i zmiany zachodzące w procesach generujących te dane. Dotyczy to, na przykład, danych pochodzących z ubezpieczalni, dużych supermarketów, danych opisujących zachowania posiadaczy kart kredytowych, klientów banku.

11. Wnioski i uwagi końcowe

Ostatnie lata pokazały olbrzymią przydatność praktyczną aplikacji eksploracji danych [1,2,3,4]. Znajdują one zastosowanie w takich dziedzinach jak: ubezpieczenia, telekomunikacja, marketing, badania naukowe, medycyna, kryminalistyka. Zasadnicza dyskusja, która się toczy obecnie w środowisku osób zajmujących się eksploracją danych dotyczy odpowiedzi na pytanie, czy eksploracja danych jest, czy powinna być, naturalnym rozszerzeniem funkcjonalności systemów baz danych, czy jest raczej dziedziną aplikacji rozwijanych dla poszczególnych zastosowań. Wydaje się, że obecnie zaczyna przeważać pogląd, iż eksploracja danych, przynajmniej w zakresie związanym z definiowaniem zapytań i generacją reguł, powinna stanowić rozszerzenie funkcjonalności systemów zarządzania bazami danych, czy raczej „systemów zarządzania bazami wiedzy". W ostatnim czasie zaproponowano szereg rozszerzeń standardu SQL umożliwiających wyszukiwanie reguł asocjacyjnych i klasyfikacyjnych, zaproponowano nowe struktury danych ułatwiających przechowywanie i wyszukiwanie reguł, nowe typy indeksów, itd., [5,6,7,8]. Są to elementy systemowe rozszerzające funkcjonalność SZBD.
Eksploracja danych jest dziedziną interdyscyplinarną. Jej dalszy rozwój wymaga połączenia technik i metod wypracowanych w różnych dziedzinach nauki: statystyce, grafice komputerowej, technologii baz danych, systemów równoległych, teorii optymalizacji, programowania matematycznego, uczenia maszynowego, itd. Część osób porównuje aktualny stan rozwoju tej dziedziny do stanu, w jakim systemy baz danych znajdowały się na początku swojej drogi. Brakuje nam standardu języka, w którym użytkownicy mogliby definiować swoje „zapytania", mechanizmów optymalizacji wykonywania takich zapytań, zarządzania efektywnym, współbieżnym wykonywaniem zapytań, narzędzi do budowy aplikacji, itd. Olbrzymie bazy danych przypominają dzisiaj czasami wielkie grobowce pełne danych, ale bez życia. Eksploracja danych jest próbą, być może kolejną, wniesienia do tych grobowców trochę światła.

Literatura.

[1] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, E. Simoudis, Industrial applications of data mining and knowledge discovery, Communicatiosn of ACM, 39, 11, 1996.

[2] U. Fayyad, G. Piatetsky-Shapiro, [eds.], Advances in Knowledge Discovery and Data Mining, MIT Press, 1996

[3] U. Fayyad, D. Haussler, P. Stolorz, Mining science data, Communications of ACM, 39, 11, 1996.

[4] T. Imieliński, H. Manilla, A Database Perspective on Knowledge Discovery, Communications of ACM, 39, 11, 1996.

[5] R. Meo, G. Psaila, S. Ceri, A new SQL-like operator for mining association rules, Proc. of 22nd Int. Conf. VLDB, 1996.

[6] T. Morzy, M. Zakrzewicz, SQL-like languages for database mining, Proc. Int. Conf. on Advances in Databases and Information Systems, 1997.

[7] T. Morzy, M. Zakrzewicz, Group Bitmap Index; A Strucure for Association Rules retrieval, Proc. Of 4th Int. Conf. On Knowledge Discovery and Data Mining, AAAI Press, 1998.

[8] T. Imieliński, A. Virmani, Association rules, and what's next? – Towards second generation data mining systems, Proc. Int. Conf. On Advances in Databases and Information Systems, 1998.

[9] R. Agrawal, T. Imieliński, A. Swami, Mining Associations between Sets of Items in Massive Databases, SIGMOD-93, Washington, 1993.

[10] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proc. of 20th Int. Conf. VLDB, 1994.

[11] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, Fast Discovery of Association Rules, in [2].

[12] R. Srikant, R. Agrawal, Mining Generalized Association Rules, Proc. of 21st Int. Conf. VLDB, 1995.

[13] J. Han, Y. Fu, Doscovery of Multiple-Level Association Rules from Large Databases, Proc. of 21st Int. Conf. VLDB, 1995.

[14] R. Srikant, R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables, SIGMOD-96, Montreal, 1996.

[15] R. J. Miller, Y. Yang, Association Rules over Interval Data, SIGMOD-97, Tucson, 1997.

[16] R. Agrawal, R. Srikant, Mining Sequential Patterns, Proc. of 11th Int. Conf. on Data Engineering, Taipei, 1995.

[17] R. Srikant, R. Agrawal, Mining Sequential Patterns: Generalization and performance Improvements, EDBT-96, Avignon, 1996.

[18] H. Mannila, H. Toivonen, A. I. Verkamo, Discovering Frequent Episodes in Sequences, KDD-95, Montreal, 1995.

[19] R. Agrawal, C. Faloutsos, A. Swami, Efficient Similarity Search in Sequence Databases, FODO-93, Chicago, 1993.

[20] C. Faloutsos, M. Ranganthan, Y. Manolopoulos, Fast Subsequence Matching in Time-Series Databases, SIGMOD-94, 1994.

[21] M. Mehta, R. Agrawal, J. Rissanen, SLIQ: A Fast Scalable Classifier for data Mining, EDBT-96, Avignon, 1996.

[22] S. M. Weiss, C. A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expwert systems, Morgan Kaufman, 1991.

[23] A. K. Jain, R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[24] T. Zhang, R. Ramakrishan, M. Livny, BIRCH: An Efficient Data Clustering Method For Very Large Databases, SIGMOD-96, Montreal, 1996.

[25] T. Morzy, M. Zakrzewicz, M. Wojciechowski, Pattern-Oriented Hierarchical Clustering, Proc. 3rd Int. Conf. on Advances in Databases and Information Systems, 1999.