Zastosowania i trendy rozwoju metod odkrywania wiedzy

Marzena Kryszkiewicz

Instytut Informatyki, Politechnika Warszawska
Nowowiejska 15/19, 00-665 Warszawa
mkr@ii.pw.edu.pl

Abstrakt. Raport wprowadza w metodologię odkrywania wiedzy z położeniem szczególnego nacisku na techniki eksploracji danych. Przedstawiono wybrane zastosowania reprezentatywnych metod odkrywania wiedzy i jej użycia takich, jak odkrywanie asocjacji, wzorców sekwencyjnych i reguł epizodycznych, klasyfikacja, predykcja, grupowanie pojęciowe i zbiory przybliżone. Omówiono także zastosowanie technik eksploracji danych do zwiększenia stopnia automatyzacji procesu projektowania bazy danych. Zilustrowano możliwości użycia różnych podejść do odkrywania wiedzy w telekomunikacji, biomedycynie, finansach, sprzedaży i analizie logów Webowych. Zarysowano nowe trendy w badaniach nad odkrywaniem wiedzy.

1 Wprowadzenie

Szacuje się, że ilość informacji w świecie podwaja się co dwadzieścia miesięcy. Wiele systemów informacyjnych (naukowych, medycznych, rządowych, biznesowych) generuje i magazynuje olbrzymie ilości danych. Współczesnie bazy danych często osiągają wielkości rzędu giga, tera, a nawet peta bajtów. Dane naukowe reprezentują obserwacje dotyczące badanych zjawisk - mogą to być np. olbrzymie kolekcje zdjęć. Znany NASA'owski System Obserwacji Ziemi (EOSDIS) gromadzi dane rzędu 10^{15} bajtów rocznie (zdjęcia z obserwacji satelitarnych) [12]. Dane gromadzone przez firmy dotyczą krytycznych rynków, konkurentów, klientów itd. Bazy tego rodzaju mogą stanowić olbrzymi potencjał wartościowej wiedzy, jednakże ich ogrom przekracza ludzkie możliwości analizowania i odkrywania wzorców. Jest to przyczyną gwałtownego rozwoju dziedziny eksploracji danych, której jednym z głównych celów jest opracowanie metod i technik umożliwiających ekstrakcję wartościowej wiedzy z bardzo dużych baz danych.

Badania te cieszą się dużym zainteresowaniem zarówno środowisk naukowych, jak i przemysłowych. W szczególności, obserwuje się wzrost zainteresowania operatorów telefonii komórkowej stosowaniem metod odkrywania wiedzy z ich obszernych zasobów informacyjnych [10, 14-16, 18, 21, 44-45, 48-49].

Raport wprowadza w zagadnienie odkrywania wiedzy ze szczególnym uwzględnieniem metodologii zaoferowanych w ostatnich latach w dziedzinie eksploracji danych. Pierwszą część raportu poświęcamy zadaniom i metodom stosowanym w odkrywaniu wiedzy ilustrując ich użycie w obszarze telekomunikacyjnym, finansowym, biomedycznym, sprzedaży i marketingu. W drugiej części przedstawiamy trendy badawcze w odkrywaniu wiedzy.

2 Metody w odkrywaniu wiedzy i ich zastosowania

Zadanie odkrywania wartościowej wiedzy z bardzo dużych baz danych, hurtowni danych i innych informacyjnych repozytoriów jest znane jako *eksploracja danych* (ang. data mining). Eksploracja danych jest dziedziną ukierunkowaną na zastosowania. Problemy badawcze, które podejmuje się w tej dziedzinie, są umotywowane specyfiką dostępnych zbiorów danych dotyczących świata rzeczywistego [25].

Korzenie eksploracji danych sięgają statystyki i sztucznej inteligencji, w tym uczenia się maszyn, sieci neuronowych, wnioskowania przez indukcję, zbiorów przybliżonych i algorytmów ewolucyjnych. W przeciwieństwie do statystyki i sztucznej inteligencji, eksploracja danych odwołuje się także do rozwiązań technologii baz danych, np. poprzez umiejętne wykorzystanie opracowanych tam metod indeksowania danych oraz metod optymalizacji wykonywania zapytań.

Jakkolwiek wiele metod odkrywania wiedzy zaoferowano w sztucznej inteligencji, większość z tych metod zawodzi, gdy próbuje się je stosować do dużych baz danych. Zazwyczaj metody sztucznej inteligencji stosuje się do zasobów danych zawierających nie więcej niż kilkanaście tysięcy rekordów.

Nowe podejście do analizy danych zainicjowano w [23, 51]. Jego najistotniejszym wyróżnikiem jest to, że jest ukierunkowane na opracowywanie efektywnych metod odkrywania wiedzy z bardzo dużych zasobów danych. W nowym podejściu, odkrywanie wiedzy jest postrzegane jako process składający sie z kilku etapów, z których eksploracja danych jest etapem centralnym. *Odkrywanie wiedzy w dużych bazach danych* (knowledge discovery in databases - KDD) obejmuje czyszczenie danych, ich integrację, selekcję, transformację, eksplorację, ocenę odkrytej wiedzy i jej prezentację [18]. Nowy nurt w odkrywaniu wiedzy rozwija się gwałtownie od początku lat 90-tych ubiegłego stulecia. Intensywne badania podjęto zarówno w dziedzinie odkrywania wzorców z hurtowni danych, klasycznych baz danych [23, 44-46], jak i z zasobów tekstowych, włączając w to badania nad eksploracją WWW [15, 24] oraz wielowymiarową analizę danych [49]. Pozycja [18] stanowi cenne podsumowanie uzyskanego dorobku. W niniejszym opracowaniu przedstawiamy reprezentatywne przykłady zastosowań metod opracowanych w tych badaniach oraz wybranych metod przejętych ze sztucznej inteligencji.

2.1. Zbiory częste i reguły asocjacyjne

Pojęcie zbioru częstego i reguł asocjacyjnych wprowadzono w [1] na przykładzie bazy danych przechowującej informacje o transakcjach dokonywanych w hipemarkecie. Zbiór produktów jest definiowany jako częsty, jeśli częstość jego występowania w bazie transakcji przekracza zadaną wartość progową. Zatem, częste zbiory produktów, to produkty, które klienci powszechnie jednocześnie nabywają. Reguła asocjacyjna natomiast ma wyrażać fakt, że zakup pewnego zbioru produktów przez klienta z dużym prawdopodobieństwem implikuje jednoczesny zakup jeszcze innych produktów w tej samej transakcji. Informacje o zbiorach częstych i regułach asocjacyjnych są wykorzystywane na wiele sposobów, począwszy od wspierania podejmowania decyzji o rozmieszczeniu produktów w poszczególnych działach i położeniu produktów na półkach sklepowych, do decyzji o tym, do których regionów,

które produkty wysyłać oraz jakie promocje kierować do poszczególnych grup klientów.

Zastosowanie metod eksploracji danych do systemów informacyjnych w bankach lub u operatorów telekomunikacji może skutkować pozyskaniem wiedzy o zachowaniach ich klientów. Przykładowa odkryta reguła mogłaby brzmieć:

W przypadku 25% klientów, jeśli klient prowadzi firmę, to wybiera plan lokaty/taryfy biznesowej z prawdopodobieństwem 95%.

Zbiory częste i reguły asocjacyjne znajdują interesujące zastosowanie w biomedycynie, analizie DNA i chemii [18, 41-42]. Stosuje się je np. do odkrywania genów często współwystępujących u osób cierpiących na tę samą chorobę. Wiedza tego typu może posłużyć do identyfikacji choroby u pacjenta, a nawet do określania jej stadium.

Kolejnym często rozważanym zadaniem eksploracji jest wykrywanie zbiorów częstych i reguł asocjacyjnych z logów Web. Interesującą odmianą tego typu zależności, którą przedstawiono w [24] jest wyszukiwanie częstych ciągłych podgrafów połączeń. Informują one o tym, że w licznych sesjach pewne strony są odwiedzane w określonej kolejności. Uzyskana wiedza o podgrafach częstych może być wykorzystana do usprawnienia połączeń sieciowych. Np. Webmaster stwierdzając, że istnieją podgrafy częste, których węzły (odpowiadajace odwiedzanym stronom) są nieosiągalne z innych podgrafów częstych, może podjąć decyzję o stworzeniu nowego połączenia.

W [24] przedstawiono także interesującą analizę wiązań w związkach rakotwórczych i nierakotwórczych o podobnym składzie chemicznym opartą na metodach odkrywania zbiorów częstych i reguł asocjacyjnych. Autorzy odkryli zależności informujące, które fragmenty wiązań w związkach chemicznych mogą być odpowiedzialne za powstawanie i rozwój raka.

Zbiory częste i reguły asocjacyjne można odkrywać także w niekompletnych bazach danych [26-27, 38].

2.2. Zależności czasowe

Wzorce sekwencyjne są asocjacjami uzwględniającymi upływ czasu. Są to takie uporządkowane w czasie podsekwencje zarejestrowanych zdarzeń/transakcji, które występują u licznej grupy badanych obiektów (klientów). Ten rodzaj eksploracji danych budzi wysokie zainteresowanie operatorów usług telekomunikacyjnych. Poniżej zamieszczamy przykład wzorca sekwencyjnego, który może być odkryty z telekomunikacyjnych zasobów danych o klientach:

65% klientów zamieszkałych w Warszawie, po zakupieniu uslugi "trzy najczęściej używane numery", zmienia plan taryfowy w przeciągu 2 miesięcy na „aktywny", a w przeciągu następnych 2 miesięcy zażąda usługi WAP.

Aplikacje wzorców sekwencyjnych na potrzeby dostawców usług telekomunikacyjnych opisano w [56].

Reguły epizodyczne podobnie jak wzorce sekwencyjne są zależnościami czasowymi. Tym razem jednak w bazie danych przechowuje się jedną sekwencję zdarzeń, a nie wiele. Z takich danych można wydobyć np. następującą regułę:

```
IF (alarm połączeniowy) AND NEXT (nieudane połączenie),
THEN alarm o wysokim współczynniku uszkodzenia [5] [60]
zaufanie [90%], częstość [151/168].
```

Reguła ta oznacza, że w 90% przypadków, jeśli w przeciągu 5 sek. wystąpił najpierw alarm połączeniowy, a następnie zarejestrowano nieudane połączenie, to w przeciagu 60 sek. wystąpił alarm o wysokim współczynniku uszkodzenia. Wszystkie 3 zdarzenia wystapiły razem 151 razy, a 2 zdarzenia z części IF reguły wystapiły 168 razy. Ponieważ w regule bierzemy pod uwagę kolejność zajścia zdarzeń w jej poprzedniku (a w ogólności również w jej następniku), to taką regułę nazywamy szeregową. Innym rodzajem reguł epizodycznych są reguły równoległe, które są wspierane także przez sekwencje zdarzeń, które zaszły w innej kolejności niż kolejność implikowana przez zapis zdarzeń w regule: Poniżej zamieszczamy przykład reguły równoległej:

```
IF (alarm połączeniowy) AND (nieudane połączenie),
THEN alarm o wysokim współczynniku uszkodzenia [5] [60]
zaufanie [90%], częstość [151/168].
```

Reguła ta oznacza, że w 90% przypadków, jeśli w przeciągu 5 sek. wystąpił alarm połączeniowy i nieudane połączenie (w dowolnej kolejności), to w przeciagu 60 sek. od wystąpienia pierwszego z wcześniejszych zdarzeń, wystąpił alarm o wysokim współczynniku uszkodzenia.

Odkrywanie reguł epizodycznych doczekało się komercyjnego zastosowania – w wyniku współpracy 4 fińskich firm telekomunikacyjnych (Nokia Telecommunications, Helsinki Telephone Corp. HPY, Radiolinja and Tampere Telephone Corp., the Technology Development Centre of Finland (Tekes)) z Uniwersytetem w Helsinkach powstał system TASA (Telecommunication Alarm Sequence Analyzer) wykorzystywany przez te firmy do odkrywania wiedzy o alarmach telekomunikacyjnych [18, 21].

Reguły epizodyczne znalazły także nietrywialne zastosowanie w przewidywanu okresów suszy z kilkumiesięcznym wyprzedzeniem. Projekty badawcze poświęcone odkrywaniu tego typu reguł są realizowane przez University of Nebraska (NSF Digital Government Grant No. EIA-0091530 i NSF EPSCOR Grant No. EPS-0091900) [17, 19-20]. Wynikiem tych projektów ma być implementacja systemu Geospatial decision support system (GDSS) – do lepszego i szybszego zapobiegania skutkom suszy. Znaczenie tych badań jest bardzo duże, co ilustrują następujące oszacowania: Średnie koszty roczne strat spowodawanych suszą w USA szacuje się na $6-8 bilionów (the Federal Emergency Management Agency), a w samym 1988r. poniesiono straty w wys. $40 bilionów (the National Climatic Data Center, Asheville NC). Ostatnio przeprowadzone badania dowodzą, np. że [20]:

- Nietypowe wartości wskaźników klimatycznych Pacyfiku i Oceanu Atlantyckiego są powiązane z długoterminowymi warunkami sprzyjającymi pojawieniu się suszy w Clay Center, Nebraska USA (środek USA).
- Globalne warunki klimatyczne mogą być wykorzystane do identyfikacji lokalnych okresów suszy na kilka miesięcy przed ich wystąpieniem.

2.3. Predykcja

Przyjmijmy, że dane są opisane przez zbiór atrybutów numerycznych. Wartość pewnych atrybutów może zależeć od wartości innych atrybutów. Pierwsze z tych atrybutów nazwiemy wejściowymi, podczas gdy drugie nazwiemy wyjściowymi. Celem zadania predykcji jest poprawne odgadnięcie-wyznaczenie (nieznanych) wartości atrybutów wyjściowych w oparciu o (znane) wartości atrybutów wejściowych. Aby wykonać operację predykcji, niezbędne jest wcześniejsze zbudowanie pewnego modelu predykcyjnego na podstawie znanego zbioru danych (lub jego próbki), dla której znane są zarówno wartości atrybutów wejściowych, jak i wyjściowych. Model predykcyjny realizuje pewną funkcję wyrażającą zależność pomiędzy atrybutami wejściowymi i wyjściowymi. Modele takie można budować stosując szereg technik począwszy od standardowej statystycznej (wielokrotnej) regresji liniowej, a skończywszy na sieciach neuronowych. Podczas gdy regresja explicite buduje formułę do obliczania wartości wyjściowej na podstawie wartości wejściowych, sieć neuronowa może być postrzegana jako czarna skrzynka symulująca (zwykle złożoną) funkcję nieliniową [8, 11]. Wadą sieci neuronowych jest to, że nie wyjaśniają zależności pomiędzy wartościami wejściowymi i wyjściowymi. Jednakże sieci neuronowe są uważane za bardzo skuteczny model predykcyjny ponieważ, w przeciwieństwie do regresji, są w stanie modelować każdą zależność pomiędzy wartościami wejściowymi i wyjściowymi. Chociaż sam process budowania tego modelu jest bardzo czasochłonny, uzyskana sieć neuronowa bardzo szybko wyznacza wartości atrybutów wyjściowych.

W telekomunikacji predykcję wykorzystuje się np. do przewidywania wielkości ruchu wynikającego z realizowanych połączeń w nowych komórkach, dostosowywania planów rozwoju sieci komórkowej, przewidywania obciążeń procesorów na podstawie liczby poszczególnych zdarzeń zachodzących w ustalonym kwancie czasu.

2.4. Klasyfikacja

Cel klasyfikacji jest zbliżony do celu predykcji. Klasyfikacja i predykcja różnią się jednakże tym, że celem predykcji jest wyznaczenie ciągłej numerycznej wartości atrybutu, podczas gdy klasyfikacja odnosi się do obiektów opisywanych atrybutami wejściowymi dowolnego typu i skutkuje przypisaniem obiektów do predefiniowanych klas. Aby nowe obiekty mogły byc sklasyfikowane, niezbędne jest wcześniejsze zbudowanie modelu klasyfikacji (*klasyfikatora*) na podstawie znanego zbioru obiektów (lub jego próbki), których przynależność do klas jest znana. Klasyfikator jest zwykle budowany w postaci drzewa decyzyjnego (lub zbioru takich drzew), ewentualnie zbioru reguł decyzyjnych. Spotyka się także klasyfikatory wykorzystujące sieci neuronowe. W przeciwieństwie do metod sztucznej inteligencji i statystycznych, niedawno zaproponowane metody eksploracji danych budują klasyfikatory z bardzo dużych zasobów danych. W szczególności, w [43, 52] zapropowano metody budowy klasyfikatorów w postaci drzew z dużych danych relacyjnych, a w [4] przedstawiono skuteczne metody budowy klasyfikatorów opartych na regułach asocjacyjnych. Nową koncepcją klasyfikatora jest klasyfikator oparty na wzorach wyłaniających się [39-41]. Nietypową cechą tego klasyfikatora jest to, że klasyfikację wykonuje także w oparciu o wzorce, które bardzo rzadko

obserwuje się w pełnym zbiorze danych. Wzorzec jest uznany za wyłaniający się, jeśli częstość jego występowania w jednej klasie jest kilkadziesiąt lub więcej razy wyższa niż w pozostałym zbiorze obiektów. Klasyfikator oparty na wzorcach wyłaniających się jest budowany z wykorzystaniem nowych technik odkrywania zbiorów częstych i posiada doskonałe własności klasyfikacyjne [39-41]. W szczególności, w przeciwieństwie, do wielu innych technik, skutecznie radzi sobie z problemem klasyfikacji obiektów należących do klas o wyjątkowo małej liczności.

W telekomunikacji klasyfikacji używa się, np. do określania nowych klientów jako należących do jednej z następujących klas: bardzo aktywny, aktywny, bierny. Wiedzę tę wykorzystuje się do umiejętnego prowadzenia kampanii reklamowych, ograniczając jej zasięg do kręgu właściwych adresatów. Tego typu wiedza może znacząco zredukować koszty marketingowe [25].

Innym typowym zastosowaniem klasyfikacji jest przewidywanie, którzy klienci zrezygnują z usług operatora. Zbudowawszy klasyfikator na podstawie dostępnych charakterystyk klientów i informacji o tym, czy zmienili operatora, czy też nie, można będzie przewidywać zamiary nowo pozyskanych klientów i odpowiednio reagować zanim klient podejmie decyzję o rezygnacji z usług. Umiejętność przewidzenia zachowania klienta jest bardzo istotna, gdyż powszechnie wiadomo, że łatwiej jest utrzymać klienta, niż pozyskać nowego.

Klasyfikatory są bardzo przydatne w bankowości, np. do wspierania decyzji o przyznaniu lub nie przyznaniu kredytu ubiegającemu się o niego petentowi.

W [4] wykazano, że klasyfikatory oparte na regułach asocjacyjnych dobrze nadają się do kategoryzacji tekstów. Kolejnym zastowaniem byłoby wykorzystanie takich klasyfikatorów do automatycznego rozdzielania wiadomości elektronicznych na prywatne, służbowe itd.

2.5. Grupowanie pojęciowe

Polega na przydzieleniu obiektów do grup (zwanych także klastrami lub segmentami) w taki sposób, że obiekty w jednej grupie są do siebie bardziej podobne niż do jakiegokolwiek obiektu przydzielonego do innej grupy. Metody grupowania pojęciowego używają różnych miar odległości wyrażających (nie)podobieństwo obiektów. Im mniejsza odległość pomiędzy obiektami, tym obiekty uznawane są za bardziej podobne. Wyznaczanie obiektów podobnych jest typową operacją w grupowaniu pojęciowym. Nowe metody grupowania pojęciowego proponowane w eksploracji danych postulują wykorzystanie do wyszukiwania obiektów podobnych indeksów wielowymiarowych [7, 13, 42] opracowanych na potrzeby wykonywania zapytań w bazach danych i hurtowaniach danych. Użycie tych technik znacząco redukuje złożoność problemu grupowania pojęciowego.

W projektowaniu metod grupowania pojęciowego ważnym zadaniem jest opracowanie miary odległości, która dobrze wyraża semantykę podobieństwa rozpatrywanych obiektów. Umiejętne dobranie miary odległości pozwoli podzielić obiekty na grupy, które ekspert będzie w stanie sensownie zinterpretować. Przypisawszy znaczenie grupom, ekspert będzie mógł je traktować jako automatycznie znalezione klasy decyzyjne i wykorzystać do zbudowania klasyfikatora. Np. grupowanie zastosowane do danych o klientach firmy telekomunikacyjnej może skutkować znalezieniem trzech grup traktowanych jako klienci: bardzo aktywni, aktywni i bierni.

Inną aplikacją metod grupowania pojęciowego jest odkrywanie wyjątków i anomalii, co z kolei może być stosowane do wykrywania fałszerstw finansowych lub włamań do systemów komputerowych (poprzez odkrywanie nietypowych charakterystyk/zachowań klientów) [8].

2.6. Zbiory przybliżone

Teoria zbiorów przybliżonych została zaproponowana jako narzędzie do analizy i wnioskowania w kontekście niedokładnej, niepewnej lub rozmytej wiedzy [50]. Metodologia ta pozwala odkrywać interesujące zależności, reguły decyzyjne i analizować sytuacje konfliktowe. Przykładowe praktyczne zastosowania teorii zbiorów przybliżonych obejmują dziedzinę telekomunikacji, medycyny, algorytmów sterowania, oceny jakości działania pilota samolotu, farmakologii, analizy wibracji, syntezy układów logicznych, przetwarzania obrazu, giełdy itd. [50, 53]. Zbiory przybliżone są postrzegane jako skuteczne niestatystyczne narzędzie do analizy danych, w tym danych niepełnych [28, 37].

Podobnie jak w większości metod eksploracji danych, wnioskowanie w tym podejściu jest oparte jedynie na dostępnych zasobach danych. Jedną z jej głównych cech jest to, że umożliwia ekstrakcję minimalnych zbiorów atrybutów (zwanych reduktami), które gwarantują jakość klasyfikacji nie gorszą niż cały zbiorów atrybutów. Ostatnio zaproponowano nową metodę w dziedzinie eksploracji danych, która pozwala efektywnie ekstrahować redukty nawet z dużych zasobów danych [33]. Dlatego zbiory przybliżone mogą być używane na etapie transformacji danych do automatycznej ekstrakcji atrybutów najbardziej stosownych do tworzenia klasyfikatorów.

Obiekty w teorii zbiorów przybliżonych są postrzegane jako nierozróżnialne jeśli mają ten sam opis w systemie. Może to być przyczyną niepewności - dwa lub więcej obiektów identycznie opisanych w systemie może należeć do różnych klas (pojęć). Takie pojęcia, jakkolwiek rozmyte, mogą być zdefiniowane w sposób przybliżony za pomocą pary zbiorów dokładnych: przybliżenia dolnego i przybliżenia górnego. Przybliżeniem dolnym pojęcia jest zbiór obiektów, które z pewnością do niego należą; natomiast przybliżeniem górnym pojęcia jest zbiór obiektów, które być może należą do tego pojęcia. Na bazie tych dwóch pojęć można generować reguły pewne i możliwe. Reguły pewne w sposób jednoznaczny klasyfikują obiekty w systemie, podczas gdy reguła możliwa klasyfikuje obiekt jako element pojęcia z pewnym prawdopodobieństwem wyznaczanym na podstawie początkowego zbioru danych. Reguły wyznaczane na podstawie teorii zbiorów przybliżonych stanowią bardzo specyficzny podzbiór reguł asocjacyjnych. Żadna reguła wyznaczona na podstawie teorii zbiorów przybliżonych nie będzie pokrywała żadnego obiektu spoza przybliżenia górnego wskazywanej przez nią klasy decyzyjnej.

Inną dziedziną, do której teoria zbiorów przybliżonych szczególnie dobrze się nadaje jest analiza konfliktów. Podejście do analizy konfliktów oparte na teorii zbiorów przybliżonych oferuje głębszy wgląd w strukturę konfliktów i umożliwia analizę związków pomiędzy koalicjami oraz kwestiami poddanymi pod dyskusję. Można wykorzystać to podejście do wspierania i uzasadniania podejmowania decyzji w obecności konfliktów [50].

2.7. Zależności funkcyjne i zależności przybliżone między atrybutami

Ustalenie zależności funkcyjnych jest jednym z niezbędnych, podstawowych zadań w procesie projektowania bazy danych. Stwierdzenie, że zbiór atrybutów *X* funkcyjnie wyznacza zbiór atrybutów *Y*, oznacza, że dla każdego możliwego zestawu wartości dla atrybutów *X* istnieje dokładnie jeden zestaw wartości dla atrybutów *Y*. Innymi słowy, *X* wyznacza funkcyjnie *Y*, jeśli każde dwa rekordy, które „zgadzają się" na atrybutach *X*, „zgadzają się" również na atrybutach *Y*. Np. wartość atrybutu pesel w sposób jednoznaczny wyznacza imię, nazwisko i datę urodzenia każdego obywatela Polski. Dotychczasowe systemy wspomagające projektowanie baz danych wymagały, aby to projektant formułował zależności funkcyjne. Próby automatycznego odkrywania zależności funkcyjnych napotykały na szereg problemów. W przypadku, gdy danych było mało, mogły nie być wystarczająco reprezentatywne i w rezultacie można było wydobyć zarówno prawdziwe zależności funkcyjne, jak i fałszywe. Gdy danych było dużo, pojawiał się problem braku skalowalnych efektywnych metod ich odkrywania. Występowanie przekłamań lub anomalii w danych również mogło skutecznie uniemożliwić odkrycie faktycznych zależności funkcyjnych. Ten ostatni problem usiłowano zniwelować wprowadzając pojęcie zależności przybliżonej. Zależność pomiędzy dwoma zbiorami atrybutów uznaje się za przybliżoną, jeśli procent rekordów nie spełniających zależności funkcyjnej pomiędzy tymi zbiorami nie przekracza zadanej wartości progowej. Przykładem zależności przybliżonej jest zależność pomiędzy atrybutem imię a atrybutem płeć - na ogół imię jednoznacznie determinuje płeć, chociaż zdarzają się nieliczne przypadki, kiedy to samo imię może posiadać zarówno kobieta, jak i mężczyzna. W ostatnich latach zaproponowano efektywny, skalowalny algorytm TANE w [22] do wyznaczania zarówno zależności funkcyjnych, jak i zależności przybliżonych. W algorytmie tym umiejętnie wykorzystuje się techniki odkrywania zbiorów częstych oraz rezultaty teorii zbiorów przybliżonych. Algortm ten jest istotnym osiągnięciem na drodze do dalszej automatyzacji procesu projektowania baz danych.

3 Trendy w odkrywaniu wiedzy

Różnorodność danych i zadań odkrywania wiedzy, oraz wielość rozwijanych technik eksploracji danych stawia szereg wyzwań badawczych w odkrywaniu wiedzy. Należą do nich: projektowanie języków eksploracji danych, rozwój efektywnych i wydajnych metod i systemów, konstrukcja interaktywnych i zintegrowanych platform umożliwiających eksplorację danych, wdrażanie technik eksploracji danych do rozwiązywania poważnych problemów aplikacyjnych. Poniżej przedstawiamy pewne trendy w odkrywaniu wiedzy, które odzwierciedlają próby odpowiedzi na te wyzwania.

- *Rozwój i standaryzacja języków eksploracji danych*. Aby ograniczyć esktrakcję informacji z baz danych do informacji, która jest nam w danym momencie przydatna, opracowano języki zapytań, które pozwalają specyfikować własności poszukiwanej informacji. Np., możemy chcieć poznać wyłącznie nazwiska studentów o średniej ocen równej 5. Podobnie, dysponujac zasobem danych, z którego można odkrywać wiedzę w różnej postaci (np. reguł, grup pojęciowych,

wzorców sekwencyjnych), warto mieć narzędzie do specyfikowania własności poszukiwanej wiedzy. Poniżej przedstawiamy przykład specyfikacji w języku M-SQL [23] własności wiedzy, którą chcemy odkryć w postaci reguł asocjacyjnych:

```
SELECT FROM MINE(Tabela) Reguły
WHERE R.ANTECEDENT >= {kategoria = *}
AND R.CONSEQUENT = {częstość_skarg = ”wysoka”}
AND R.SUPPORT > 10% AND R.CONFIDENCE >= 65%;
```

Zapytanie to wyraża zamiar użytkownika (pracownika firmy telekomunikacyjnej), który chce poznać przyczyny skarg zarejestrowanych w *Tabeli* występujących z prawdopodobieństwem co najmniej 65% w co najmniej 10% przypadków. Jako rezultat wykonania tego zapytania system mógłby zwrócić następujące reguły:

```
IF kategoria IN {”niski poziom sygnału”} AND od = {”aktywacja”},
THEN częstość_skarg = ”wysoka” [wsparcie = 12%, zaufanie = 95%];

IF kategoria IN {”echo”} AND od = {”miesiąc”},
THEN częstość skarg = ”wysoka” [wsparcie = 15%, zaufanie = 70%].
```

Reguły te nie muszą być produktem końcowym. Na przykład użytkownik może zażądać wyszukania wszystkich rekordów, które potwierdzają (ang. support) wskazaną regułę lub jej zaprzeczają (ang. violate). Reguły mogą być użyte do sklasyfikowania nieznanych obiektów itd. Proponowane języki eksploracji danych dopuszczają odkrywanie także innych typów wiedzy, takich jak klasyfikatory, reguły epizodyczne, wzorce sekwencyjne, grupy pojęciowe, wyjątki [18, 23, 44-46]. Prowadzenie dalszych intensywnych prac w zakresie rozwoju i standaryzacji języków eksploracji danych jest niezbędne.

- *Opracowanie skalowalnych metod eksploracji danych z nałożonymi więzami*. Posiadając możliwość specyfikowania własności pożądanej wiedzy, użytkownik oczekuje, że system odkrywania wiedzy zwróci tylko ten fragment wiedzy, który spełnia nałożone więzy. Na etapie eksploracji danych można oczywiście wydobyć wiedzę zarówno interesującą użytkownika, jak i nie interesującą, a dopiero potem odfiltrować wiedzę interesującą. Zwykle jednak wiedza interesująca stanowi maleńką cząstkę pełnej wiedzy, którą wydobywa się w czasie eksploracji danych. Istotne jest opracowanie metod, które już bezpośrednio w samym procesie eksploracji danych będą wykorzystywać ograniczenia nałożone przez użytkownika do szybkiej eliminacji wiedzy nie interesującej. Często już na etapie samej inicjalizacji procesu eksploracji danych można znacząco ograniczyć przestrzeń rozwiązań. Inne podejście sprowadza się do redukcji (logicznej lub fizycznej) bazy danych, z której mają być odkryte wzorce. Jeśli np. użytkownik chce odkrywać silne reguły zawierające więcej niż 4 pozycje w następniku, to takich reguł na pewno nie odkryje z transakcji zawierających co najwyżej 4 pozycje. Dlatego wszystkie transakcje z co najwyżej 4 pozycjami mogą być usunięte z bazy danych bez obawy, że odkrywanie ze zredukowanej bazy doprowadzi do utraty pewnych reguł poszukiwanych przez użytkownika. Podobnie, ograniczenia nałożone na postać odkrywanej wiedzy mogą implikować możliwość redukcji atrybutów w eksplorowanym zbiorze danych. Efektywne

wykorzystanie informacji o więzach w procesie eksploracji danych może przyczynić się do znacznego skrócenia czasu wykonania tego etapu [55].

- *Opracowanie zwięzłych reprezentacji różnych typów wiedzy i wykorzystanie ich do efektywnego odkrywania wiedzy z nałożonymi więzami*. Jedną z głównych wad odkrywania wiedzy z dużych lub silnie skorelowanych zasobów danych jest olbrzymia liczba uzyskiwanych wyników. Ostatnie badania pokazują, że można ekstrahować tylko najbardziej istotną i zwięzłą wiedzę, która umożliwia wyprowadzenie pozostałej wiedzy na żądanie bez dostępu do bazy danych. Modele reprezentacji zbiorów częstych, reguł asocjacyjnych i epizodycznych zaproponowane w literaturze [5-6, 9, 29-32, 34-36] są nawet o kilka rzędów wielkości bardziej zwięzłe niż cała wiedza, jaką można pozyskać z danych. Podobnie, proces ekstrakcji tych modeli jest o kilka rzędów wielkości szybszy niż ekstrakcja całej wyprowadzalnej wiedzy. Nowym wyzwaniem jest opracowanie modeli reprezentacji wiedzy interesującej (ew. dostosowanie istniejących reprezentacji), przy czym poprzez wiedzę interesującą rozumiemy wiedzę posiadającą własności wyspecyfikowane przez użytkownika.
- *Integracja eksploracji danych z systemami baz danych, hurtowaniami danych i systemami internetowych baz danych*. W ostatnich latach obserwuje się ciekawą wymianę osiągnięć pomiędzy technologią baz danych a metodami eksploracji danych. Np. nowe metody grupowania pojęciowego korzystają z najnowszych osiągnięć w zakresie indeksowania danych, a zaadaptowane metody odkrywania zbiorów częstych są wykorzystane do odkrywania zależności funkcyjnych [22] i realizacji pewnych klasycznych zapytań bazo-danowych [18]. Sukcesy w realizacji wcześniej wymienionych trendów dodatkowo będą sprzyjać integracji eksploracji danych z różnego typu systemami baz danych. Przewiduje się powstanie komercyjnych indukcyjnych systemów baz danych jako jednolitej platformy wyposażonej zarówno w funkcjonalność systemu bazy danych, jak i funkcjonalność hurtowni danych oraz funkcjonalność systemu eksploracji danych.
- *Rozwój specjalizowanych (problemowo-zorientowanych) systemów eksploracji danych*. Trend ten wynika z doświadczeń w zakresie wykorzystania eksploracji danych w praktyce. W początku lat 90-tych rozważano przede wszystkim zastosowania biznesowe. Obecnie zastosowania różnicują się i obejmują biomedycynę, analizę finansową oraz telekomunikację. Często dziedzina problemu jest wysoce specyficzna i wymaga bardzo żmudnej i skomplikowanej wstępnej obróbki danych pod kierunkiem eksperta w tej dziedzinie. Samo zadanie odkrywania pożytecznej wiedzy z dużych zasobów danych także często różni się mniej lub bardziej od zadań rozważanych w literaturze. Wymaga to dostosowywania znanych rozwiązań do warunków specyficznych dla dziedziny problemu lub opracowywania całkiem nowych rozwiązań. Nie jest możliwe stworzenie systemu, który byłby w stanie uwzględniać specyfikę wszystkich możliwych zastosowań.
- *Rozwój metod eksploracji danych złożonych*. Eksploracja danych złożonych jest ważnym wyzwaniem badawczym. Jakkolwiek poczyniono już pewien postęp w technikach eksploracji danych przestrzennych, multimedialnych, tekstowych, internetowych, sekwencji i szeregów czasowych, to jednak ich złożoność jest zbyt duża, aby móc je stosować w praktyce. Pootrzebne jest dalsze prowadzenie intensywnych badań nad eksploracją tych typów danych.

- *Rozwój wizualnych technik eksploracji danych*. Wizualizacja ma ogromne znaczenie zarówno w fazie początkowej odkrywania wiedzy, jak i końcowej. W fazie początkowej ułatwia zrozumienie charakteru i właściwości danych, które mają być eksplorowane, co wpływa na sposób przygotowania danych (np. poprzez wybór stosownej metody dyskretyzacji danych). Wizualizacja jest także bardzo przydatna do prezentowania rezultatów uzyskanych w procesie eksploracji danych. Umożliwia zlokalizowanie wiedzy istotnej w zbiorze rezultatów wygenerowanych w sposób automatyczny w procesie eksploracji danych. Systematyczne studia nad wizualizacją eksploracji danych mogą poszerzyć jej zastosowanie do efektywnego sterowania całym procesem odkrywania wiedzy [54].
- *Ochrona prywatności i bezpieczeństwa informacji*. Zagadnienie to jest niezmiernie ważne. Udostępniając fragment wiedzy odkrytej w danych warto mieć świadomość tego, co jeszcze z tego fragmentu można wyprowadzić. Jak się okazuje, niekiedy z pozornie nieistotnych fragmentów wiedzy można nawet bez sięgania do bazy danych wyprowadzić informację pewną o pierwszorzędnym znaczeniu [31], która nie powinna być ujawniona osobom nieuprawnionym. Innym zadaniem, które także muszą adresować bezpieczne techniki eksploracji danych jest zapewnienie, że z pozyskanej wiedzy osoba niepowołana nie będzie w stanie wyprowadzić chronionej prawem informacji o charakterze prywatnym [3]. Ostatnio zaczyna pojawiać się coraz więcej prac badawczych poświęconych zagadnieniu ochrony prywatności i zapewnieniu bezpieczeństwa dzielonej informacji.

4 Konkluzje

Odkrywanie wiedzy jest dziedziną zorientowaną na zastosowania, w której problemy badawcze są często motywowane dostępnością i specyfiką zasobów informacyjnych o świecie rzeczywistym. Proces odkrywania wiedzy często wymaga współpracy analityków, ekspertów w dziedzinie rozważanego problemu, informatyków i statystyków. Innowacyjne metody eksploracji danych są coraz powszechniej używane. Najwięcej udanych zastosowań eksploracji danych można zaobserwować w dziedzinie telekomunikacji, biomedycynie i analizie DNA, finansach, sprzedaży i marketingu.

Jednym z najważniejszych trendów w odkrywaniu wiedzy jest stworzenie indukcyjnych systemów baz danych, które efektywnie i wydajnie będą realizować zadania eksploracji danych formułowane w stosownych językach specyfikacji pożądanej wiedzy. Systemy te powinny zapewnić ochronę danych prywatnych oraz bezpieczny dostęp do wiedzy.

Bibliografia

[1] Agrawal R., Imielinski T., Swami A.: Mining Associations Rules between Sets of Items in Large Databases. In: Proc. of the ACM SIGMOD Conference on Management of Data, Washington, USA (1993) 207-216

[2] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast Discovery of Association Rules. In: Advances in KDD. AAAI, Menlo Park, California (1996) 307-328

[3] Agrawal R., Srikant R.: Privacy-Preserving Data Mining. In: Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, 2000. SIGMOD Record, Vol. 29, No. 2 (2000) 439–450

[4] Antonie M.L., Zaïane O.R.: Text Document Categorization by Term Association. ICDM 2002: 19-26 Bayardo R.J., Agrawal R.: Mining the Most Interesting Rules. In: Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA, USA, 1999. ACM (1999) 145-154

[5] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets. CL (2000) 972-986

[6] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining Frequent Patterns with Counting Inference. ACM SIGKDD Explorations, Vol. 2(2). (2000) 66-75

[7] Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. SIGMOD Conference 1990: 322-331

[8] Berry M. J. A., Linoff G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons (1997)

[9] Bykowski, A., Rigotti, C.: A Condensed Representation to Find Frequent Patterns. In: Proc. of the 12th ACM SIGACT-SIGMOD-SIGART PODS' 01 (2001)

[10] Daszczuk W., Gawrysiak P., Gerszberg T., Kryszkiewicz M.: Mieścicki J., Muraszkiewicz M., Okoniewski M., Rybiński H., Traczyk T., Walczak Z.: Data Mining for Technical Operation of Telecommunications Companies: a Case Study. In: Proc. of 4th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, USA (2000) 64-69

[11] Duda R.O., Hart P.E., Stork D.G.: Pattern Classification. . John Wiley & Sons (2001)

[12] EOSDIS – NASA's Earth Observation System, http://www.earth.nasa.gov/

[13] Ester M., Kriegel H.-P., Sander J., Xu X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996: 226-231

[14] Gajek, M.; Gancarz, Ł.; Muraszkiewicz, M.; Stueckemann, W.; Rybiński, H.: "Współpraca Instytutu Informatyki z operatorami telekomunikacyjnymi" , Przegląd i Wiadomości Telekomunikacyjne, nr 10, 2001

[15] Gawrysiak P., Okoniewski M.: Knowledge Discovery in the Internet. Archiwum Informatyki PAN (2001)

[16] Gawrysiak P., Okoniewski M., Rybiński H.: Regression - Yet Another Clustering Method. In: Advances in Soft Computing. Proc. of International Intelligent Information Systems Conference. Springer (2001)

[17] Goddard S., Harms S.K., Reichenbach S.E., Tadesse T., Waltman W.J.: Geospatial decision support for drought risk management. CACM 46(1): 35-37 (2003)

[18] Han J., Kamber M.: Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers (2000)

[19] Harms S.K., Deogun J.S., Saquer J., Tadesse T.: Discovering Representative Episodal Association Rules from Event Sequences Using Frequent Closed Episode Sets and Event Constraints. ICDM 2001: 603-606

[20] Harms S.K., Deogun J.S., Tadesse T.: Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. ISMIS 2002: 432-441

[21] Hätönen K., Klemettinen M., Mannila H., Ronkainen P., Toivonen H.: Knowledge Discovery from Network Alarm Databases. In Proc. of the Twelfth International Conference on Data Engineering (ICDE), New Orleans, Louisiana, 1996. IEEE Computer Society (1996) 115-122

[22] Huhtala Y., Kärkkäinen J., Porkka P., Toivonen H.: TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies. The Computer Journal 42(2): 100-111 (1999)

[23] Imielinski T., Mannila H.: A Database Perspective on Knowledge Discovery. Communications of the ACM, Vol. 39, No. 11 (1996) 58-64

[24] Inokuchi A., Washio T., Motoda H.: Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. Machine Learning 50(3): 321-354 (2003)

[25] Kleinberg J. M., Papadimitriou C., Raghavan P.: A microeconomic view of data mining. Data Mining and Knowledge Discovery, 2:311-324 (1998)

[26] Kryszkiewicz M.: Association Rules in Incomplete Databases. In: Methodologies for Knowledge Discovery and Data Mining. Proc. of Third Pacific-Asia Conference (PAKDD). Beijing, China, 1999. Lecture Notes in Computer Science, Vol. 1574. Springer (1999) 84-93

[27] Kryszkiewicz M.: Probabilistic Approach to Association Rules in Incomplete Databases. In: Proc. of Web-Age Information Management Conference (WAIM), Shanghai, China, 2000. Lecture Notes in Computer Science, Vol. 1846. Springer-Verlag (2000) 133-138

[28] Kryszkiewicz M.: Rough Set Approach to Rules Generation from Incomplete Information Systems. In: The Encyclopedia of Computer Science and Technology. Marcel Dekker, Inc., New York, Vol. 44 (2001) 319–346

[29] Kryszkiewicz, M.: Closed Set based Discovery of Representative Association Rules. In: Proc. of IDA '01. Springer (2001) 350-359

[30] Kryszkiewicz, M.: Concise Representation of Frequent Patterns based on Disjunction-Free Generators. In: Proc. of ICDM '01. IEEE (2001) 305-312

[31] Kryszkiewicz M.: Inferring Knowledge from Frequent Patterns. In: Computing in an Imperfect World. Proc. of First International Conference Soft-Ware, Belfast, Northern Ireland, 2002. Lecture Notes in Computer Science, Vol. 2311. Springer (2002) 247-262

[32] Kryszkiewicz M.: Closed Set Based Discovery of Maximal Covering Rules. In: Proc. of Ninth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Annecy, France (2002) 299-306

[33] Kryszkiewicz M.: Scalable Methods of Discovering Rough Sets Reducts, manuskrypt

[34] Kryszkiewicz M., Gajek M.: Concise Representation of Frequent Patterns based on Generalized Disjunction-Free Generators. In: Advances in Knowledge Discovery and Data Mining. Proc. of 6th Pacific-Asia Conference (PAKDD). Taipei, Taiwan, 2002. Lecture Notes in Computer Science, Vol. 2336. Springer (2002) 159-171

[35] Kryszkiewicz M., Gajek M.: Why to Apply Generalized Disjunction-Free Generators Representation of Frequent Patterns? In: Foundations of Intelligent Systems. Proc. of 13th International Symposium (ISMIS) Lyon, France, 2002. Lecture Notes in Artificial Intelligence, Vol. 2366. Springer-Verlag (2002) 383-392

[36] Kryszkiewicz M., Rybiński H., Gajek, M.: Concise Representations of Frequent Patterns. Dataless Transitions between Concise Representations of Frequent Patterns. Przyjęto do druku w Intelligent Information System Journal (JIIS)

[37] Kryszkiewicz M., Rybiński H.: Incompleteness Aspects in Rough Set Approach. In: Proc. of International Joint Conference of Information Sciences, Raleigh, North Carolina, USA, Vol. 2 (1998) 371-374

[38] Kryszkiewicz M., Rybiński H.: Legitimate Approach to Association Rules under Incompleteness. In: Proc. of Foundations of Intelligent Systems. Proc. of 12th International Symposium (ISMIS), Charlotte, USA, 2000. Lecture Notes in Artificial Intelligence, Vol. 1932. Springer-Verlag (2000) 505-514

[39] Li J., Dong G., Ramamohanarao K.: Instance-Based Classification by Emerging Patterns. PKDD 2000: 191-200

[40] Li J., Dong G., Ramamohanarao K.: Making Use of the Most Expressive Jumping Emerging Patterns for Classification. Knowledge and Information Systems 3(2): 131-145 (2001)

[41] Li J., Wong L.: Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns . Bioinformatics 18(10): 1406-1407 (2002)

[42] Markl V., Bayer R.: Processing relational OLAP queries with UB-Trees and multidimensional hierarchical clustering. DMDW 2000: 1

[43] Mehta M., Agrawal R., Rissanen J.: SLIQ: A Fast Scalable Classifier for Data Mining. EDBT 1996: 18-32

[44] Meo R., Psaila G., Ceri S.: A New SQL-like Operator for Mining Association Rules. In: Proc. of 22th International Conference on Very Large Data Bases (VLDB). Mumbai, India, 1996. Morgan Kaufmann (1996) 122–133

[45] Microsoft Corporation. OLE DB for Data Mining and Knowledge Discovery, Ver. 0.9. In: http://www.microsoft.com/data/oledb/dm.html (2000)

[46] Morzy T., Zakrzewicz M.: SQL-Like Language for Database Mining. In: Proc. of the First East-European Symposium on Advances in Databases and Information Systems (ADBIS). St. Petersburg, Russia, 1997. Nevsky Dialect (1997) 311–317

[47] Muraszkiewicz M.: Data Mining at a Glance. Proc. of Int'l Conf. Tempus PHARE, Gdańsk, Poland (1999)

[48] Muraszkiewicz M.: Eksploracja danych dla telekomunikacji (Data Mining for Telecommunications). In: Systemy informatyczne w dobie Internetu. Proc. of VI Konferencja PLOUG (2000)

[49] Okoniewski M., Gancarz Ł., Gawrysiak P.: Mining Multi-Dimensional Quantitative Associations. In: Proc. of the 14th International Conference on Applications of Prolog (INAP), Tokyo, Japan, 2001. Prolog Association of Japan (2001) 265-274

[50] Pawlak Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Vol. 9 (1991)

[51] Piatetsky-Shapiro G.: Discovery, Analysis and Presentation of Strong Rules. In: Piatetsky-Shapiro G., Frawley W. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press, Menlo Park, CA (1991) 229-248

[52] Shafer J.C., Agrawal R., Mehta M.: SPRINT: A Scalable Parallel Classifier for Data Mining. VLDB 1996: 544-555

[53] Slowinski R.: Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Vol .11 (1992)

[54] Westphal C., Blaxton T.: Data Mining Soultions: Methods and Tools for Solving Real-World Problems. John Wiley & Sons (1998)

[55] Wojciechowski M., Zakrzewicz M.: Dataset Filtering Techniques in Constraint-Based Frequent Pattern Mining. Pattern Detection and Discovery 2002: 77-91

[56] Wu P.H., Peng W.C., Chen M.S.: Mining Sequential Alarm Patterns in a Telecommunication Database. Databases in Telecommunications (2001) 37-51