Marta CHODYKA, Zofia LUBAŃSKA, Tomasz GRUDNIEWSKI

ODKRYWANIE WIEDZY W BAZACH DANYCH

Streszczenie

Rozwój technologii automatycznego gromadzenia i przechowywania informacji jest wynikiem nagłego wzrostu rozmiarów przechowywanych danych i eksplozją danych. Nowe możliwości daje ciągły wzrost mocy obliczeniowej i pojemności pamięci, jednak posiadanie danych nie jest równoznaczne z posiadaniem wiedzy zawartej w tych danych. Dla podejmowania lepszych decyzji i wyciągnięcia wniosków konieczna staje się zaawansowana analiza dużych wolumenów danych. Techniki eksploracji danych pozwalają na znajdowanie nieznanych zależności pomiędzy danymi, co przyczynia się do wspomagania podejmowania decyzji.

Celem artykułu jest prezentacja rozwiązań związanych z odkrywaniem wiedzy w bazach danych, poprzez omówienie technik eksploracji danych oraz reprezentacji odkrywanej wiedzy z jej praktycznym zastosowaniem.

WSTĘP

Wzrost rozmiarów przechowywanych danych i eksplozja danych przyczyniły się do rozwoju technologii automatycznego gromadzenia i przechowywania informacji we współczesnym świecie. Nowe możliwości daje także ciągły wzrost mocy obliczeniowej i pojemności pamięci. Należy mieć jednak na uwadze, iż posiadanie danych nie jest równoznaczne z posiadaniem wiedzy zawartej w tych danych. Dla podejmowania lepszych decyzji i wyciągnięcia wniosków konieczna staje się zaawansowana analiza dużych wolumenów danych.

Istniejące komputerowe systemy wspomagania decyzji bazują na wiedzy ekspertów i na analizie zawartości bazy danych. Przykładem są tu systemy typu OLAP (ang. *Online Analitycal Processing*). Środowisko OLAP charakteryzuje się stosunkowo nielicznymi, ale za to złożonymi transakcjami odczytu. Miarą efektywności jest czas odpowiedzi. Powszechnie wykorzystywane jest w technikach związanych z eksploracją danych. Praca w środowisku analitycznym zakłada potrzebę przygotowania przez ekspertów hipotez, których poprawność jest weryfikowana przy pomocy narzędzi OLAP. Możliwe jest usprawnienie tego procesu poprzez zautomatyzowanie odkrywania wiedzy w bazach danych (ang. *Knowledge Discovery in Databases*) – generowanie i automatyczne przygotowanie hipotez. Wymaga to wówczas oceny i akceptacji odkrytej wiedzy z wykorzystaniem wskaźników statystycznych. Dodatkowo techniki eksploracji danych pozwalają na znajdowanie nieznanych zależności pomiędzy danymi, co przyczynia się do jeszcze efektywniejszego wspomagania podejmowania decyzji.

Dostępne są na rynku zintegrowane środowiska programowe umożliwiające odkrywanie wiedzy w najbardziej popularnych systemach zarządzania bazami danych. Zaliczyć do nich można m.in. Data Mining Suite – Information Discovery, MineSet – Silicon Graphics, Intelligent Miner – IBM, Modeler – Integral Solutions.

1. PROCES ODKRYWANIA WIEDZY

Odkrywanie wiedzy (ang. *Knowledge Discovery*) stanowi proces poszukiwania nowych, użytecznych potencjalnie regularności w danych [7]. Celem tego procesu jest przejście od początkowych danych do zbioru wzorców, które mogą być w następnym kroku wykorzystane w procesie wspomagania podejmowania decyzji. Na proces odkrywania wiedzy składają się następujące etapy:

a) analiza i poznanie dziedziny zastosowania, identyfikacja dostępnej wiedzy i celów użytkownika,
b) integracja danych z różnych źródeł, czego celem jest integracja danych z różnych heterogenicznych i rozproszonych źródeł danych w jeden zintegrowany zbiór danych,
c) tworzenie/selekcja danych istotnych z punktu widzenia procesu analizy danych,
d) czyszczenie i wstępne przetwarzanie danych, czyli usunięcie niepełnych, niepoprawnych lub nieistotnych danych ze zbioru eksplorowanych danych,
e) transformacja danych (przekształcenie i redukcja danych) wyselekcjonowanych do postaci wymaganej przez metody eksploracji danych,
f) wybór zadania lub metody eksploracji danych,
g) wybór algorytmu eksploracji danych,
h) eksploracja danych (ang. *Data Mining*), czyli automatyczne odkrywanie nietrywialnych danych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców schematów, podobieństw lub trendów w dużych repozytoriach danych, np. hurtowniach danych, jest to najistotniejszy etap odkrywania wiedzy,
i) interpretacja, analiza i ocena odkrytej wiedzy, wizualizacja odkrytych wzorców, czyli identyfikacja interesujących wzorców oraz ich wizualizacja w taki sposób, aby umożliwić użytkownikowi ich interpretację i zrozumienie,
j) przygotowanie wiedzy do użycia.

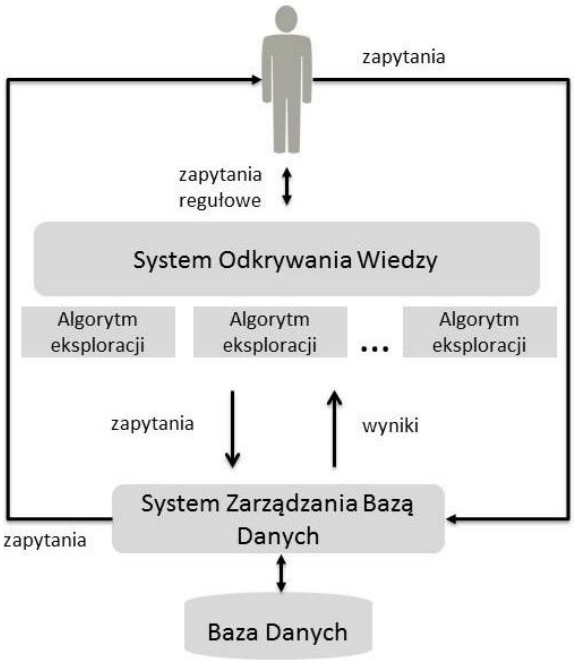
2. TECHNIKI EKSPLORACJI DANYCH

Poszukiwana wiedza może przyjmować wiele postaci [1], m. in. reguły asocjacyjne, drzewa, reguły klasyfikacyjne, czy trendy i zależności czasowe. Wiąże się to z określonymi zadaniami eksploracji danych. Metody te zostały szeroko omówione w literaturze [6]. W zależności od zmiennej wyjściowej rozróżnia się klasyfikację, w przypadku zmiennej jakościowej lub regresję, w przypadku zmiennej liczbowej [2,5]. Klasyfikacja to znajdowanie sposobu odwzorowywania danych w zbiór predefiniowanych klas, wiąże się z predykcją przynależności do klas decyzyjnych tj. kategorii w oparciu o zbiór danych treningowych. Do technik klasyfikacji zaliczyć można techniki statystyczne, drzewa decyzyjne, reguły decyzyjne, sieci neuronowe. Regresja natomiast to predykcja wartości funkcji rzeczywistej w oparciu o zbiór danych treningowych. Do technik regresji zaliczyć można sieci neuronowe i statystyczne metody regresji. Do zadań eksploracji danych należą też grupowanie danych, czyli znajdowanie skończonych zbiorów klas obiektów posiadających podobne cechy, odkrywanie wzorców sekwencji -

poszukiwanie zależności pomiędzy występowaniem określonych zdarzeń w czasie, odkrywanie zależności funkcyjnych - poszukiwanie wzorów najlepiej wyrażających zależności występujące pomiędzy atrybutami o wartościach liczbowych. Jest to w pewnym sensie uogólnienie regresji na dowolną liczbę atrybutów zależnych z dodatkowym wymogiem, aby zależność była wyrażona za pomocą formuły algebraicznej [3]. Do znajdowania takich zależności wykorzystuje się metody odkrywania równań, analiza przebiegów czasowych oraz wyszukiwanie według zawartości – poszukiwanie wzorców podobnych do podanego wzorca. Zadanie to jest wykonywane najczęściej na zbiorach danych zawierających teksty i obrazy. W przypadku tekstu wzorcem może być np. zbiór słów kluczowych, a w przypadku obrazów, użytkownik może posiadać przykładowy obraz, szkic lub opis obrazu, co umożliwia interaktywne przeszukiwanie bazy obrazów z wykorzystaniem deskryptorów zawartości.

3. ARCHITEKTURA SYSTEMU ODKRYWANIA WIEDZY

System odkrywania wiedzy to specjalistyczne oprogramowanie, które jest niezbędne w procesie odkrywania wiedzy. Potrzeba jego stosowania jest nieodzownie związana z rozmiarem rozwiązywanych problemów i potrzebą szybkiego dostępu do eksplorowanych danych i umożliwia składowanie odkrytej wiedzy w bazie danych. Architektura Systemu Odkrywania wiedzy przedstawiona została na Rysunku 1.

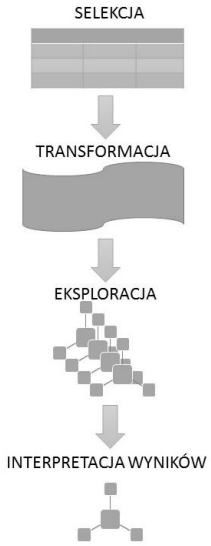


Rys. 1. *Architektura Systemu Odkrywania Wiedzy*

Użytkownik odpowiada za specyfikację zapytań, tzw. zapytań regułowych, w których określa jakich reguł poszukuje i jakie dane mają być eksplorowane w celu odkrywania reguł. W zależności od żądanego typu reguł, system odkrywania wiedzy wykorzystuje odpowiedni algorytm eksploracji danych i wysyłane jest zapytanie do systemu zarządzania baza danych w celu znalezienia tych reguł. Następnie przeprowadzana jest filtracja znalezionych reguł, w celu dopasowania kryteriów zapytania regułowego skierowanego do systemu przez użytkownika. W efekcie końcowym użytkownik otrzymuje jako wynik jego zapytania zbiór reguł. W celu komunikacji użytkownika i aplikacji z systemem, istnieje szereg języków zapytań regułowych. W dużej mierze są to rozszerzenia istniejącego standardu SQL o dodatkowe polecenia, operatory i typy danych. Przy rozszerzaniu języka SQL, wykorzystywany jest wspólny interfejs zarówno do wyszukiwania danych, jak również do generowania reguł na podstawie tych danych.

4. ZASTOSOWANIA

Funkcjonujące zintegrowane środowiska programistyczne umożliwiają odkrywanie wiedzy w popularnych systemach zarządzania bazami danych. Produkty te związane są z fazami odkrywania wiedzy przedstawionymi na schemacie (Rysunek 2).



Rys. 2. *Fazy Odkrywania Wiedzy*

Selekcja danych to wybór krotek, które mają być w kolejnej fazie eksplorowane. W czasie transformacji danych następuje konwersja typów atrybutów i dyskretyzacja wartości ciągłych [4]. Na etapie eksploracji następuje ekstrakcja wiedzy z danych, generowanie drzew decyzyjnych, reguł czy sieci neuronowych. Jest to najistotniejszy etap procesu odkrywania wiedzy, gdzie za pomocą wyboru właściwego algorytmu dla zależności danych, użytkownik otrzymuje dane w postaci formalnej. W końcowym etapie tj. interpretacji odkrytej wiedzy następuje wybór wiedzy i wizualizacja wyników.

Do zintegrowanych środowisk programowe, które umożliwiają odkrywanie wiedzy w systemach zarządzania bazami danych należą m.in. Data Mining Suite, MineSet, Intelligent Miner, Modeler.

Data Mining Suite do korzystania z baz danych wykorzystuje interfejs SQL. Przeznaczony jest do odkrywania wiedzy w bardzo dużych wolumenach danych. Odkrywanie wiedzy może być nadzorowane przez użytkownika, bądź może przebiegać automatycznie. Wspiera techniki eksploracji danych takie jak: klasyfikacja, odkrywanie charakterystyk, analiza zależności, klastering, odkrywanie odchyleń. MineSet zaś wspiera metody eksploracji takie jak: odkrywanie reguł asocjacyjnych, klasyfikacja na podstawie niepełnych danych, klasyfikacje za pomocą drzew decyzyjnych. Środowisko to dostarcza narzędzi do wizualizacji wiedzy, umożliwia animację i trójwymiarową wizualizację danych, reguł i drzew decyzyjnych. Dane mogą być pobierane bezpośrednio z systemów baz danych. Zestaw narzędzi realizujących algorytmy odkrywania klasyfikacji i asocjacji, wykrywania odchyleń, czy klastrowania zawarty jest w pakiecie Intelligent Miner, który korzysta z architektury klient – serwer. Pakiet Modeler umożliwia pobieranie danych z bazy danych, z plików tekstowych lub z arkusza kalkulacyjnego. Dane mogą być przedstawione w postaci graficznej i cały pakiet posiada interfejs programowania graficznego. Użytkownik decyduje jak dane będą pobierane, eksplorowane i jak mają być prezentowane wyniki.

Istnieje szereg narzędzi i oprogramowania do eksploracji danych i odkrywania wiedzy. Należy do nich także WEKA, Statistica, Matlab, R, Rapid Miner i inne.

Na potrzeby badań stworzona została baza danych właściwości obrazów, na której z wykorzystaniem systemu zarządzania bazą danych Oracle testowane zostały przedstawione w artykule techniki eksploracji danych. Obrazy o różnych formatach i pochodzeniu – głównie z mikroskopii skaningowej i atomowej stanowiły bogate źródło danych w odkrywaniu wiedzy. Praca na bazie danych obrazów jest całkowicie odmienna od baz tekstowych. Zastosowanie technik eksploracji również wymaga więcej pracy. W tym przypadku opis zawartości obrazów tworzony był manualnie, jednak opis właściwości obrazu tworzony był automatycznie.

PODSUMOWANIE

Zaprezentowany w artykule model odkrywania wiedzy w bazach danych i eksploracji danych staje się coraz popularniejszy.

Automatyczne gromadzenie i przechowywanie informacji jest wynikiem nagłego wzrostu rozmiarów przechowywanych danych i eksplozją danych. Należy mieć jednak na uwadze, iż posiadanie danych nie jest równoznaczne z posiadaniem wiedzy zawartej w tych danych. Dla podejmowania lepszych decyzji i wyciągnięcia wniosków konieczna staje się zaawansowana analiza dużych wolumenów danych. Techniki eksploracji danych pozwalają na znajdowanie nieznanych zależności pomiędzy danymi, co przyczynia się do wspomagania podejmowania lepszych decyzji.

Przedstawione w artykule narzędzia umożliwiają odkrywanie wiedzy w systemach zarządzania bazą danych. Do przykładowych zastosowań eksploracji danych zaliczyć można przede wszystkim: marketing, analizy finansowe, wykrywanie nieprawidłowości i anomalii, Text mining i Web minig, czy wiele innych działów.

BIBLIOGRAFIA

1. Chodyka M. *Technologia Oracle Data Mining jako współczesna metoda analizy danych,* Varia Informatyka : technologie i bezpieczeństwo / red. M. Miłosz, P. Muryjas. - Lublin : Polskie Towarzystwo Informatyczne, 2006. - s. 19-25
2. Cichosz P., *Systemy uczące się*, WNT, Warszawa, 2000.
3. Hand D., Mannila H., Smyth P., Eksploracja danych. WNT, Warszawa, 2005.
4. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*., Springer, 2001.
5. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, WNT, Warszawa, 2005.
6. Kwiatkowski W., *Metody automatycznego rozpoznawania wzorców*, 2007
7. Larose D. T., *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych* PWN, Warszawa, 2006.

KNOWLEDGE DISCOVERY IN DATABASES

Abstract

Development of technologies for automated collection and storage of information is a result of a sudden increase in the size and amount of storage data. Continuous increase of computing power and storage capacity gives new opportunities, but having the data is not synonymous with having the knowledge contained in these data. For making better decisions and drawing conclusions, it is necessary to make an advanced analysis of large volumes of storage information. Data mining techniques (called. Data Mining) allow you to find the unknown relationship between the data, which helps to support decision-making.

Autorzy:

dr inż. **Marta Chodyka –** Państwowa Szkoła Wyższa im. Papieża Jana Pawła II w Białej Podlaskiej, Wydział Nauk Ekonomicznych i Technicznych, Katedra Nauk Technicznych, Zakład Informatyki, m.chodyka@dydaktyka.pswbp.pl

mgr inż. **Zofia Lubańska** – Państwowa Szkoła Wyższa im. Papieża Jana Pawła II w Białej Podlaskiej, Wydział Nauk Ekonomicznych i Technicznych, Katedra Nauk Technicznych, Zakład Informatyki, z.lubanska@dydaktyka.pswbp.pl

doc. dr inż. **Tomasz Grudniewski –** Państwowa Szkoła Wyższa im. Papieża Jana Pawła II w Białej Podlaskiej, Wydział Nauk Ekonomicznych i Technicznych, Katedra Nauk Technicznych, Zakład Informatyki, knt@pswbp.pl