XI Konferencja PLOUG
Kościelisko
Październik 2005

Zastosowanie reguł asocjacyjnych, pakietu Oracle Data Mining for Java do analizy koszyka zakupów w aplikacjach e-commerce. Integracja ze środowiskiem Oracle JDeveloper

Krzysztof Kawa

empolis arvato

e–mail: krzysztof.kawa@empolis.com

Streszczenie
Referat ma na celu pokazanie praktycznych możliwości, jakie daje pakiet Oracle Data Mining (ODM) przy analizowaniu koszyka zakupów tzw. market basket analysis. Obok części ściśle teoretycznej przedstawiającej filozofię analizowania danych w oparciu o reguły asocjacyjne, referat zawiera drugą część będącą case study. Referat pokazuje możliwość integracji modułu ODM z aplikacjami pisanymi w języku JAVA w środowisku Oracle JDeveloper.

Informacja o autorze
Krzysztof Kawa, absolwent informatyki na Politechnice Poznańskiej w Poznaniu oraz Wydziału Biznesu Międzynarodowego na Akademii Ekonomicznej w Poznaniu. Od kilku lat pracuje w kraju i zagranicą przy projektowaniu i eksploatacji dużych systemów informatycznych typu content management oraz systemów zarządzania wiedzą.

1. Eksploracja danych i jej etapy

Definicja eksploracji danych

Eksplorację danych (ang. data mining) definiuje się jako proces automatycznego i efektywnego wykrywania nieznanych dotychczas zależności w zbiorach danych.

Pomimo faktu, iż sama eksploracja jest kluczowym etapem odkrywania wiedzy (Np. reguł asocjacji), zwykle stanowi jedynie część całego procesu. Resztę czasu zajmują pre-etapy samego przygotowania danych.

Etapy eksploracji danych

- analiza problemu (poznanie jego natury, specyfiki)
- selekcja istotnych danych / czyszczenie danych (Np. usuwanie danych znacząco odstających wartościami od innych, usuwanie pustych wartości)
- konwersja typów danych, dyskretyzacja wartości ciągłych
- eksploracja
- wizualizacja wyników eksploracji
- weryfikowanie uzyskanych wyników
- ewentualne powtórzenie eksploracji z zastosowaniem innych algorytmów/parametrów eksploracji
- zastosowanie otrzymanej wiedzy

Różne klasy eksploracji danych

- klasyfikacja
- klastrowanie
- odkrywanie podobieństw w przebiegach czasowych
- odkrywanie asocjacji
- wykrywanie zmian i odchyleń

W niniejszej pracy zajmę się bliżej tylko jedną z tych klas - regułami asocjacyjnymi. Aby nauczyć się interpretować rezultaty uzyskiwane podczas eksploracji danych przy wykorzystaniu reguł asocjacyjnych, zostanie przedstawione jak działa sam algorytm wyznaczanie reguł asocjacyjnych. W dalszej części zostanie przedstawione jak oprogramowanie firmy Oracle wspomaga wspomniane wcześniej etapy eksploracji danych a także jak przy jego pomocy zbudować system wspomagający analizowanie koszyka zakupów (ang. basket analysis). Na końcu pracy zostaną przedstawione typowe zastosowania reguł asocjacyjnych.

2. Definicje. Wyznaczanie reguł asocjacyjnych

Pomimo, iż reguły asocjacyjne mogą być stosowane do odkrywania wiedzy nie tylko w tak zwanej analizie koszykowej, w dalszej części artykułu ograniczę się jedynie do niej. Założę także, iż dane wejściowe mają postać transakcyjną – rysunek 1 i nie zawierają „błędnych" przypadków.

Nr transakcji	Nazwa kupionego towaru
1	Piwo
1	Chipsy
1	Orzeszki
2	Cukier
3	Piwo
3	Chipsy

Rys. 1. Przykład danych w postaci transakcyjnej

Oznaczenia:

- pozycje (ang. items) opisują dostępne towary I = {i1, i2, ..., im} – zbiór wszystkich towarów
- baza transakcji D = {(tid1, T1), (tid2, T2)...} gdzie tidj - unikalny identyfikator i Tj $\subset$ I – zbiór elementów występujących w jednej transakcji
- itemset: każdy podzbiór zbioru elementów I; k-itemset: – podzbiór k-elementowy

Wparcie i zaufanie, zbiór częsty

Zanim przejdę do omówienia, czym są i jak wyznacza się reguły asocjacyjne niezbędne będzie poznanie 2 definicji: wsparcia (ang. support) oraz zaufania (ang. confidence) .

Definicję wsparcia w sposób formalny możemy zapisać jako:

$$\mathbf{support(X \Rightarrow Y) = s(X \cup Y)}$$

Opisowo wsparcie możemy wyrazić jako stosunek transakcji wspierający dany zbiór do wszystkich transakcji.

Zaufanie zaś w sposób formalny możemy zapisać następująco:

$$\mathbf{confidence(X \Rightarrow Y) = s(X \cup Y) / s(X)}$$

Opisowo określa ono prawdopodobieństwo występowania implikacji, iż w transakcji będzie występował podzbiór X (poprzednik reguły) i Y (następnik reguły).

Zbiorem częstym nazywamy zbiór o wsparciu nie mniejszym od zadanego wsparcia minimalnego.

Zobaczmy to na przykładzie:

Nr trans.	Kupione towary					
1	T0		T2	T3		T5
2	T0	T1	T2	T3		
3			T2		T4	T5
4	T0	T1	T2	T3		
5	T0		T2			T5

Reguła	support	support
T0⇒T2	80 %	100 %
T0,T2⇒ T3	60 %	75 %

Rys. 2. Przykład obliczania wsparcia i zaufania

Zachodzącą regułę:

```
T0⇒T2  wsparcie 80% , zaufaniu 100%
```

możemy zinterpretować następująco: 100 % osób, którzy kupili towar T0 kupili również towar T2 a sytuacja ta zachodzi w 80 % wszystkich transakcji.

Wyznaczanie reguł asocjacyjnych

Dane wejściowe:

- zbiór pozycji I = {i1, i2, ..., im}
- baza transakcji D = {(tid1, T1), (tid2, T2)...}
- *sup_min* = minimalna wartość wsparcia i *conf_min* = minimalny stopień wiarygodności

Cel

Znaleźć wszystkie reguły asocjacyjne o wsparciu ≥ sup_min i stopniu wiarygodności ≥ conf_min.

Cel pozornie prosty wcale takim nie jest. Przeglądniecie wszystkich dostępnych kombinacji itemsetów jest w większości przypadku po prostu nie możliwe ze względu na ich ogromną ilość. Większość istniejących algorytmów działa dwu-fazowo.

- Znajdowanie zbiorów częstych o wsparciu większym niż to podane jako parametr wejściowy (*sup_min*)
- Dla każdego wyznaczonego zbioru częstego, znajdź reguły spełniające kryterium dotyczące minimalnej wartości zaufania

Algorytm zapisany w postaci pseudokodu może wyglądać następująco:

```
C1 := I; F1 := rodzina 1-elem. zbiorów częstych
for (k = 2;Fk-1≠0;k++) do
Ck := AprioriGen(Fk-1);
//generowanie nowych kandydatów
Fk := {X ∈ Ck : support(X)  >= min_sup}
end for
Wynik := F1 ∪ F2 ∪ F3... ∪ Fk;
```

3. Oracle Data Miner a reguły asocjacyjne

Przygotowanie środowiska

Postępując zgodnie z dokumentacją, po zainstalowaniu bazy danych Oracle 10g, zainstalowaniu stosownej wersji patcha, zainstalowaniu produktów z Oracle 10g Companion CD należy utworzyć w bazie danych specjalną przestrzeń tabel, specjalnego użytkownika oraz odpalić skrypt ładowania przykładowych danych.

```
SQLPLUS sys/<sys_password> as sysdba
```

Tworzenie przestrzeni tabel

```
SQL>@<ORACLE_HOME>\dm\admin\odmtbs.sql <user tablespace> <full path to table-
space file>
```

Tworzenie użytkownika

```
SQL> @<ORACLE_HOME>\dm\admin\odmuser.sql <odm_user> <odm_password> <user ta-
blespace>
```

Ładowanie przykładowych danych

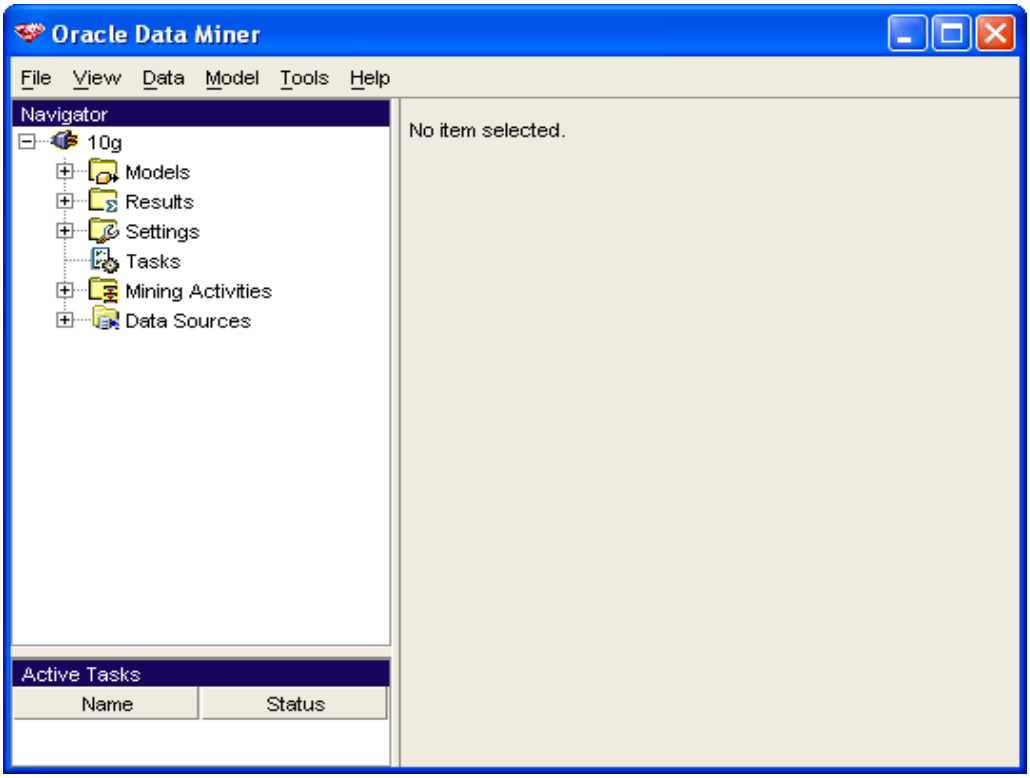
```
SQL>@<ORACLE_HOME>\dm\admin\dmuserld.sql<odm_user><odm_password> <ORACLE_HOME>
<SQLLDR_TEMP_FILE>
```

Po przygotowaniu samej bazy danych możemy przystąpić do uruchomienia Oracle Data Miner.

Program uruchamiamy wydając komendę:

```
C:\Dminer\bin>odminer.exe
```

Po wystartowaniu powinniśmy zobaczyć to, co pokazuje rysunek 3.



Rys. 3. Oracle Data Miner

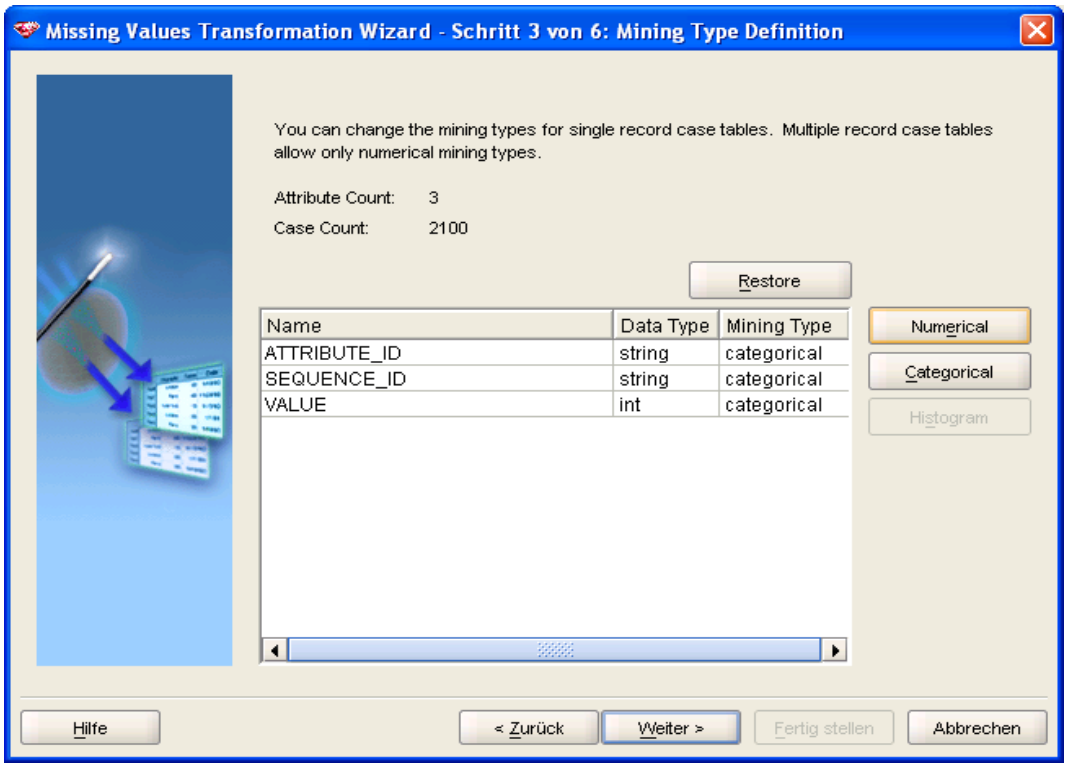
Jak Oracle Data Miner wspomaga przygotowanie danych do eksploracji

Jak zostało to wspomniane wcześniej, samo generowanie reguł asocjacyjnych stanowi jedynie stosunkowo niewielką część w procesie eksploracji (zwykle 25-30 % całego czasu). Resztę czasu pochłania przygotowanie samych danych do eksploracji (Np. ich czyszczenie) oraz przygotowanie wyników w postaci bardziej czytelnej dla większości użytkowników (wizualizacja).

Oracle Data Miner wspomaga również przygotowywanie danych, między innymi w zakresie:

- filtrowania danych
- dyskretyzacji
- normalizacji
- uzupełniania brakujących wartości

Operacje te można wykonać za pomocą dostępnych kreatorów – rysunek 4.

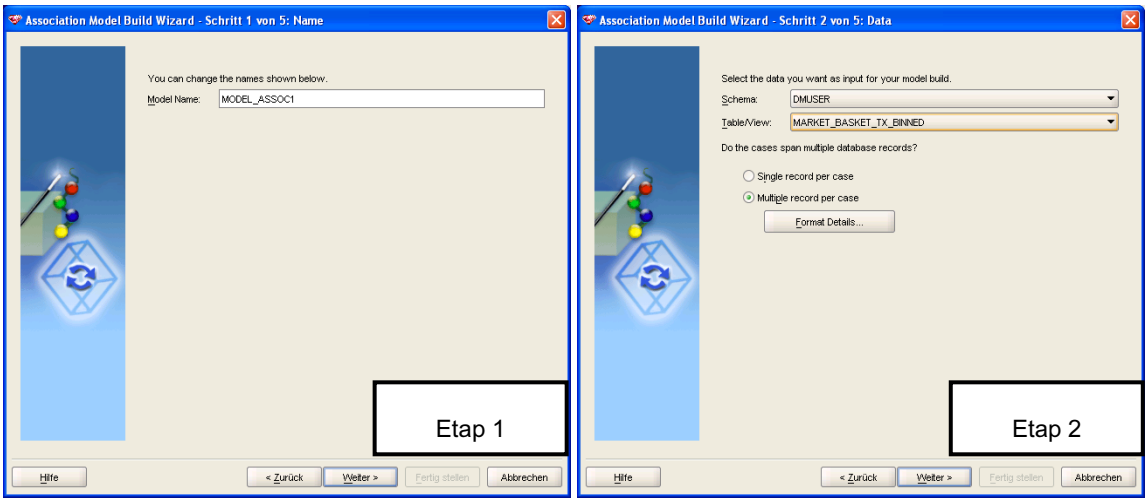


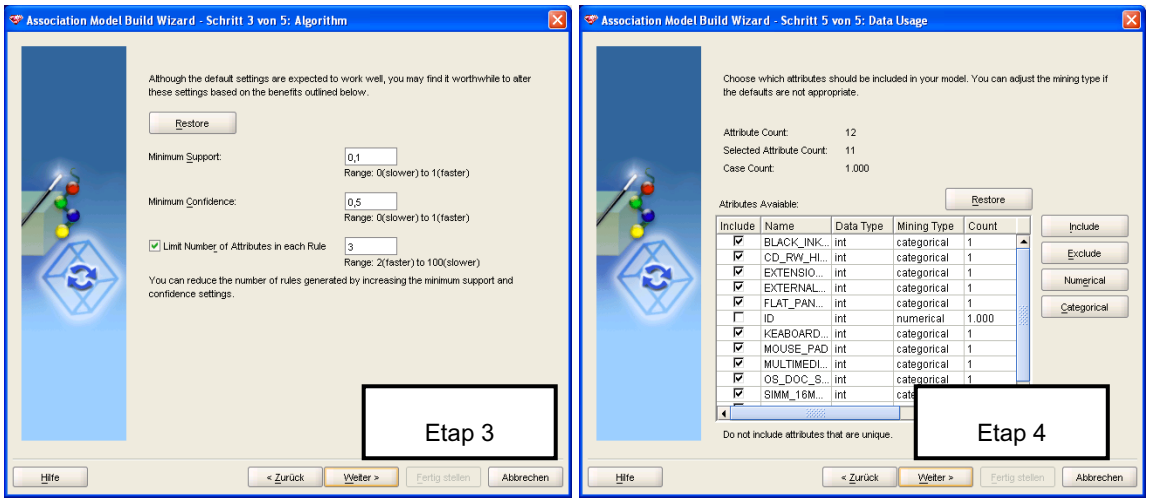
Rys. 4. Kreator przygotowywania danych (usuwanie brakujących wartości)

Generowanie reguł asocjacyjnych przy pomocy Oracle Data Miner

Przedstawię teraz jak przy pomocy Oracle Data Miner można przeprowadzić analizę koszyka zakupów.

Jako danych wejściowych użyję przykładowego schematu - danych w postaci transakcyjnej (zwartość tabeli podobna jest do tej na rysunku 1 i zawiera transakcje zakupu części komputerowych). Rysunek 5 przedstawia kolejne kroki kreatora reguł asocjacyjnych.

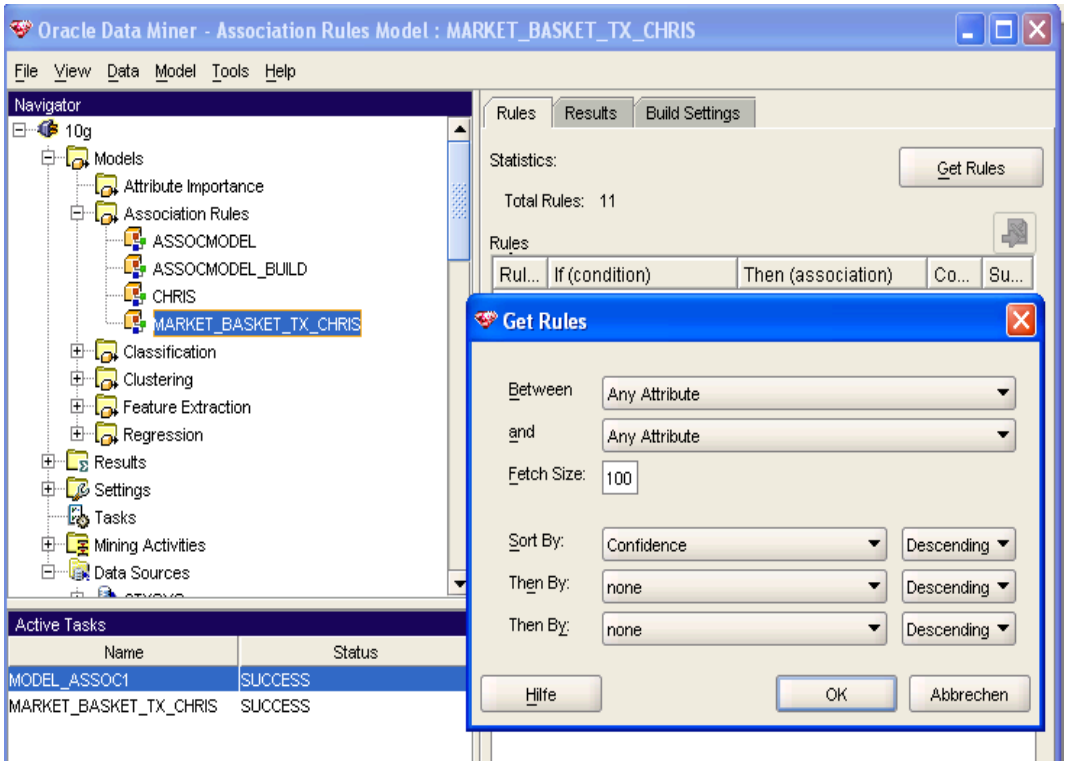




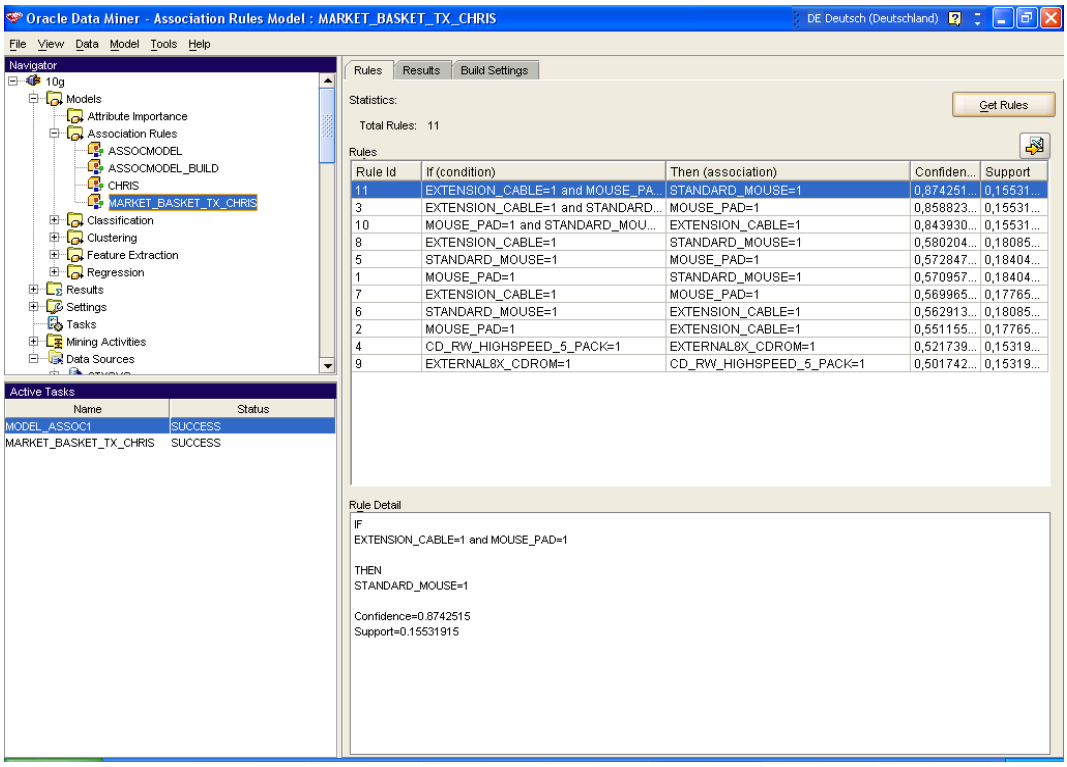
Rys. 5. Etapy generowania reguł asocjacyjnych przy pomocy Oracle Data Miner

- **Etap1** – ustalenie nazwy budowanego modelu
- **Etap2** – wybór tabeli z danym (dane w postaci transakcyjnej)
- **Etap3** – określenie parametrów algorytmu (min_sup=0,1, min_conf=0,5)
- **Etap4** – wybór atrybutów

Chcąc obejrzeć wygenerowane reguły rozwijamy w drzewie pozycje Models ->Association Rules a w niej stworzony wcześniej model. Wciskając przycisk Get Rules pojawi się dialog, w którym możemy doprecyzować jak, ile i jakie reguły zostaną pokazane.



Rys. 6. Pobieranie reguł



Rys. 7. Reguły

Wygenerowane reguły mogą być zapisane w formacie Excela lub pliku CSV.

Rys. 8. Eksport wyników

Jak można zobaczyć na rysunku 7, program w czasie eksploracji wyznaczył 11 reguł asocjacyjnych.

Przyjrzyjmy się bliżej następującej regule:

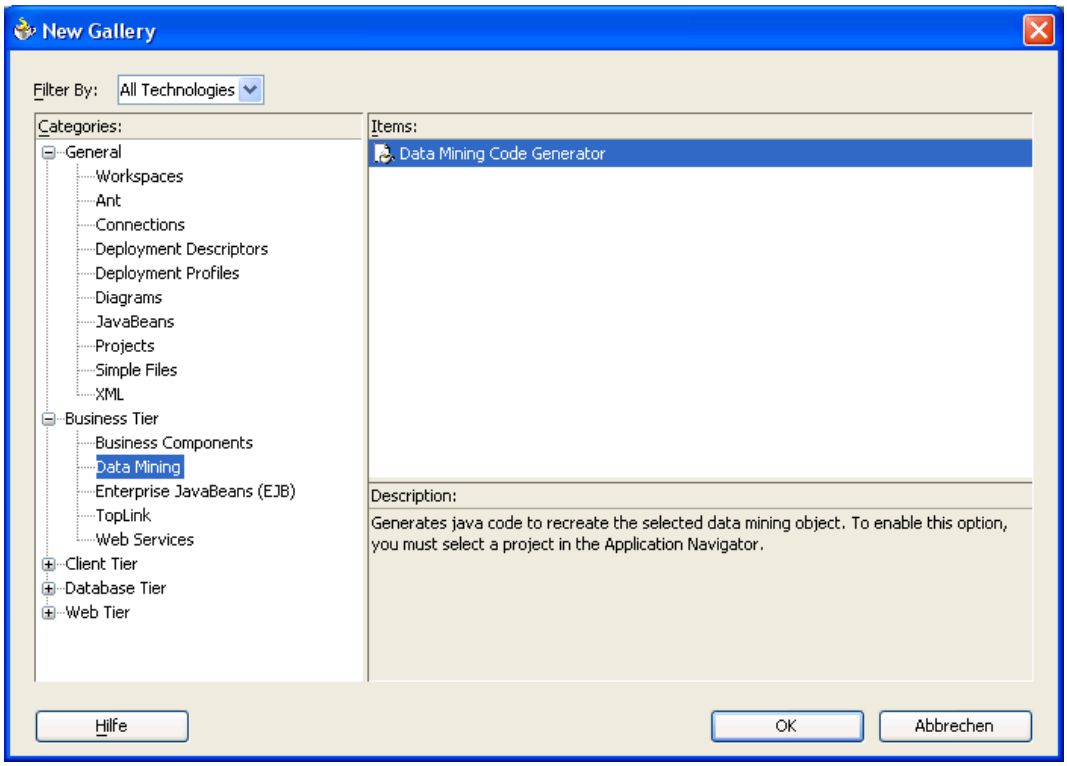
```
IF
EXTENSION_CABLE=1 and MOUSE_PAD=1
THEN
STANDARD_MOUSE=1

Confidence=0.8742515
Support=0.15531915
```

Oznacza ona, że 87 % procent osób, którzy kupili EXTENSION_CABLE oraz MOUSE_PAD kupili również STANDARD_MOUSE , a sytuacja ta zachodzi w 15 %.

Oracle Data Miner a Java. Integracja z JDeveloper 10g

Generowanie reguł asocjacyjnych możemy również uruchomić z poziomu JDevelopera 10g. Niezbędne jest zainstalowanie specjalnego rozszerzenia. Jego instalacja polega na przekopiowaniu do katalogu `JDeveloper10g\jdev\lib\ext\` plików plug-ina.

Po ich przekopiowaniu i uruchomieniu JDeveloper-a wśród dostępnych kreatorów projektu pojawi się nowy – rysunek 9.



Rys. 9. JDeveloper 10g, Data Mining Code Generator

Oracle Data Mining dostarcza API zarówno dla języka PL/SQL jak i Java. Oba interfejsy umożliwiają na komunikowanie się osobno napisanych aplikacji z modułem data mining. Pod adresem http://jcp.org/aboutJava/communityprocess/final/jsr073/ można zapoznać się dokładniej z tym co udostępnia interfejs.

4. Zastosowania reguł asocjacyjnych

Najczęściej spotykanym zastosowaniem reguł asocjacyjnych jest tzw. analiza koszykowa (ang. basket analysis). Wynikiem analizy koszykowej jest zestaw reguł opisujących zachodzące zależności między kupowanymi towarami w jednej transakcji. Przykładowa reguła może wyglądać następująco:

Przy zakupie orzeszków ziemnych w 80 % było kupowane piwo

Wiedza wynikająca z reguł może wspomagać decyzje managerów wyższego stopnia na przykład podczas:

- pozycjonowania produktów na półkach. Produkty często występujące (kupowane) w jednej transakcji umieszcza się w sklepach blisko siebie (Np. piwo, orzeszki)
- opracowywania strategii sprzedaży (obniżenie ceny tylko jednego z produktów o silnie skorelowanej sprzedaży powoduje osiągnięcie wzrostu sprzedaży obu produktów)
- opracowywania akcji marketingowych

W aplikacjach e-commerce np. sklep internetowy Amazon bardzo często spotyka się listy podpowiedzi pokazujące, iż klienci, którzy kupili produkt A zwykle też byli zainteresowani produktem B.

Reguły asocjacyjne mogą być wykorzystywane podczas analizy dowolnych danych transakcyjnych a takie można spotkać w ubezpieczeniach, bankowości, telekomunikacji, medycynie.

Bibliografia

[Mo99] Morzy T.: Eksploracja danych: problemy i rozwiązania Materiały konferencyjne PLOUG1999, Zakopane

Ci00] Cichosz P.: Systemy uczące się WNT, 2000, ISBN 83-204-2544-1

[Hu05] Nguyen Hung Son : Reguły asocjacyjne

[Or05] Oracle Data Mining, http://www.oracle.com/technology/products/bi/odm/index.html