Politechnika Poznańska
Wydział Informatyki i Zarządzania
Instytut Informatyki

Praca dyplomowa magisterska

RÓWNOLEGŁE ODKRYWANIE REGUŁ ASOCJACYJNYH ZAIMPLEMENTOWANE NA PROCESORY GRAFICZNE

inż. Tomasz Kujawa

Promotor
dr inż. Witold Andrzejwski

Poznań, 2011

Tutaj przychodzi karta pracy dyplomowej;
oryginał wstawiamy do wersji dla archiwum PP, w pozostałych kopiach wstawiamy ksero.

Spis treści

Rozdział 1

Wstęp

Proces informatyzacji przedsiębiorstw, rozpoczęty kilka dekad temu, wprowadził światową gospodarkę na nowe, dotąd nieznane tory rozwoju. Skrócenie procesu produkcyjnego, wprowadzenie kontroli komputerowych, czy też skomputeryzowanych maszyn skróciło i ułatwiło produkcję, a także zarządzanie procesami w firmach i przedsiębiorstwach. Przed ludźmi stanęły możliwości, ale także wyzwania, z którymi nigdy wcześniej nikt nie musiał sobie radzić. Zmiany, jakie nastąpiły przez ostatnie trzy dekady są nieodwracalne i zmuszają informatyków do tworzenia nowych aplikacji, które będą w stanie sprostać wymaganiom im stawianym.

Informatyzacja firm, instytucji oraz innych jednostek organizacyjnych powinna realizować dwa podstawowe cele. Z jednej strony powinna ona usprawniać pracę pojedynczego pracownika poprzez automatyzację realizowanych przez niego rutynowych zadań. Dzięki wykorzystaniu możliwości komputerów działania te powinny być wykonywane szybciej i w sposób bardziej niezawodny. Z drugiej strony celem informatyzacji jest wpływanie na działanie całych firm w wyniku wspomagania decyzji kadry zarządzającej przedsiębiorstwami. Szybka analiza bazująca na pełnej i aktualnej informacji o stanie firmy może ułatwić kadrze zarządzającej podejmowanie trafnych i szybkich decyzje o strategicznym znaczeniu dla rozwoju danego przedsiębiorstwa.

Wprowadzenie komputerów do właściwie każdej przestrzeni ludzkiego życia wpłynęło na wyprodukowanie olbrzymich ilości danych. Reprezentowane są one w sposób umożliwiający ich składowanie i przetwarzanie komputerowe przez aplikacje analityczne. W chwili obecnej ludzkość jest świadkiem eksplozji ilości danych produkowanych przez różnego rodzaju systemy komputerowe. Analiza tych danych przynieść może wymierne korzyści nie tylko w kwestiach finansowych, ale również poznawczych. Dzięki analizie danych zebranych w przeszłości możliwe jest lepsze dopasowanie planów w przyszłości - na tej podstawie planowane mogą być np. akcje marketingowe, czy też promocje w supermarketach spożywczych. Wykorzystanie wiedzy uzyskanej w ten sposób jest niezwykle szerokie i może być użyte w każdym rejonie działalności firmy.

Odkrycie zależności pomiędzy zgromadzonymi danymi bez zastosowania narzędzi informatycznych jest procesem bardzo skomplikowanym i wymagającym do realizacji dużo czasu. Przy obecnej złożoności większości systemów oraz rozmiarom danych produkowanych przez te systemy, kosz czasowy jest na tyle duży, że ręczna analiza tych danych stała się niemożliwa. Dlatego też tworzone są narzędzia umożliwiające odkrywanie prawidłowości w dużych zbiorach danych, by człowiek na tej podstawie mógł podejmować decyzje i wyciągać wnioski.

Dział informatyki, który zajmuje się odkrywaniem ukrytych dla człowieka prawidłowości i reguł w danych nazywa się eksploracją danych (ang. *Data Mining*), który jest jednym z etapów procesu *odkrywania wiedzy z baz danych* (*KDD*, ang. *Knowledge Discovery in Databases*). Proces odkrywania wiedzy w bazach danych obejmują zwykle działania bardziej złożone niż tylko eksploracja

danych. Eksploracja danych to proces odkrywania wiedzy w postaci nowych, użytecznych, poprawnych i zrozumiałych wzorców w bardzo dużych wolumenach danych [FPSSU96]. Możliwości stosowania technik eksploracji danych w praktyce, wymagają efektywnych metod przeszukiwania ogromnych plików lub baz danych. Warto przy tym wspomnieć, że tego typu technologie nie są w chwili obecnej dobrze zintegrowane z systemami zarządzania bazami danych.

Eksploracja danych (w literaturze spotkać można również określenie drążenie danych, ekstrakcja danych, pozyskiwanie wiedzy, czy też wydobywanie danych) odbywa się najczęściej w środowisku baz lFub hurtowni danych, które stanowią doskonałe źródła danych do analizy - głównie ze względu na łatwość dostępu oraz usystematyzowaną strukturę przechowywanych informacji. Ponieważ liczba odkrytych wzorców w wielu przypadkach może być bardzo duża, odkryte wzorce bardzo często zapisuje się w osobnych relacjach bazy lub hurtowni danych. Pozwala to na ich dalsze przetwarzanie w trybie off-line przez użytkowników końcowych. Pojęcie eksploracji zyskuje coraz większą popularność (również w wymiarze marketingowym) i jest wykorzystywane w wielu dziedzinach ludzkiego życia.

Jednym z najczęściej wykorzystywanych modeli wiedzy w eksploracji danych są reguły asocjacyjne. Reguła asocjacyjna ma postać $\mathbf{X} \rightarrow \mathbf{Y}$, gdzie $\mathbf{X}$ oraz $\mathbf{Y}$ są wzajemnie rozłącznymi zbiorami elementów. Przykładem reguły, która mogła zostać odkryta w badzie danych sklepu komputerowego, może być reguła postaci $komputer \wedge myszka \Rightarrow monitor$. Reguła ta prezentuje fakt, że klienci kupujący komputer oraz myszke z dużym prawdopodobieństwem kupią również monitor. W [Agr94] po raz pierwszy sformułowany został problem odkrywania reguł asocjacyjnych. Podstawą wielu algorytmów odkrywania reguł asocjacyjnych jest algorytm Apriori, zaprezentowany w [AIS93]. Dokładniejszy opis algorytmu znajduje się w rozdziale ? niniejszej pracy.

W ostatnich latach pojawiły się nowe możliwości wykorzystania współczesnych komputerów. W roku 2007 firma nVidia udostępniła programistom silnik CUDA (ang. *Compute Unified Device Architecture*), który umożliwia wykorzystanie do obliczeń *procesorów graficznych* (*GPU*, ang. *Graphics Processing Unit*). Do tej pory bardzo małe jest zainteresowanie wykorzystaniem tej technologii w procesie odkrywania wiedzy, a w szczególności znajdowania reguł asocjacyjnych. Wyniki przeprowadzonych eksperymentów pozwalają przypuszczać, że algorytm ten będzie wyraźnie szybszy od klasycznych algorytmów eksploracji danych.

1.1 Cel i zakres pracy

Celem pracy jest zaprojektowanie i zaimplementowanie algorytmu odkrywającego reguły asocjacyjne, który będzie wykorzystywał możliwośći współczesnych kart graficznych dzięki wykorzystaniu technologii CUDA. Poza tym nastąpi również porównanie zaprojektowanego i zaimplementowanego algorytmu do innych, podstawowych algorytmów odkrywania reguł asocjacyjnych. W ramach pracy dokonane zostanie również zebranie wiedzy dotyczącej algorytmów eksploracji reguł asocjacyjnych.

Rozdział 1 - wstęp.. Tutaj dalszy opis struktury pracy - zrobiony na koniec, gdy wszystko dalej będzie już znane.

Rozdział 2

Podstawy teoretyczne

W rozdziale tym przedstawiony zostanie przegląd literatury, który stanowi podstawy wiedzy na temat eksploracji danych, a w szczególności problemu odkrywania reguł asocjacyjnych w dużych zbiorach danych.

2.1 Definicje

2.1.1 Model teoretyczny

Niech $\mathbf{I} = \mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m$ będzie zbiorem binarnych atrybutów, zwanych elementami lub towarami. Niech $\mathbf{T}$ będzie zbiorem transakcji. Każda z transakcji t jest reprezentowana, jako wektor wartości binarnych, gdzie $t[k] = 1$ oznacza, że w transakcji t zakupiony został element $\mathbf{I}_k$, w przeciwny wypadku $t[k] = 0$. W bazie danych znajduje się jedna krotka odpowiadająca jednej transakcji. Niech $\mathbf{X} \subseteq \mathbf{I}$, wówczas transakcja t spełnia $\mathbf{X}$, jeżeli dla wszystkich elementów $\mathbf{I}_k \in \mathbf{X}$, $t[k] = 1$.

Definicja 1 *Reguła asocjacyjna ma postać* $\mathbf{X} \Rightarrow \mathbf{Y}$*, gdzie* $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{I}$ *oraz* $\mathbf{X} \cap \mathbf{Y} = \oslash$.

Reguła asocjacyjna zdefiniowana w Definicji 1 określa, że jeśli dany klient kupił towary ze zbioru $\mathbf{X}$, najprawdopodobniej kupi też towary ze zbioru $\mathbf{Y}$. Aby reguła asocjacyjna stanowiła interesujące źródło informacji dla analityka stosującego techniki eksploracji danych, musi ona spełniać określone warunki wyrażone za pomocą odpowiednich miar. Dwie najbardziej popularne miary jakości reguł asocjacyjnych to poziom pokrycia (
englishsupport) oraz poziom ufności (ang. *confidence*).

Definicja 2 *Poziom pokrycia dla reguły asocjacyjnej postaci* $\mathbf{X} \Rightarrow \mathbf{Y}$ *jest miarą definiowaną dla zbioru* $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ *- określa ona częstotliwość wystąpień zbioru* $\mathbf{Z}$ *w zbiorze transakcji* $\mathbf{T}$.

Oznacza to, że poziom pokrycia (*sup*) jest stosunkiem liczby transakcji zawierających elementy sumy zbiorów (czyli $|\mathbf{Z}|$, gdzie $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$) do liczby wszystkich transakcji w systemie ($|\mathbf{T}|$). Wzór 2.1 przedstawia sposó obliczania wartości *sup*.

$$sup = \frac{|\mathbf{Z}|}{|\mathbf{T}|} \tag{2.1}$$

Łatwo zauważyć, że jeśli poziom ten jest niski, to oznacza to, że nie ma jednoznacznych dowodów na łączne występowanie elementów zbioru $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$, ponieważ zbiór $\mathbf{Z}$ występuje w niewielkiej liczbie transakcji.

Poziom ufności jest miarą zdefiniowaną dla implikacji reprezentowanej przez regułę asocjacyjną [EN05].

Definicja 3 *Pioziom ufności reguły asocjacyjnej* $\mathbf{X} \Rightarrow \mathbf{Y}$ *jest równy*

$$conf = \frac{sup(\mathbf{Z})}{sup(\mathbf{X})} \tag{2.2}$$

gdzie $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$.

Ze wzoru wynika, że poziom ufności jest prawdopodobieństwem tego, że elementy tworzące zbiór $\mathbf{Y}$ zostaną kupione przez danego klienta pod warunkiem, że klient kupi elementy należące do zbioru $\mathbf{X}$.

Ważnym jest by nie mylić poziomu ufności z poziomem pokrycia. Podczas gdy ufność określa "siłę" reguły, pokrycie podkreśla jej statystyczną ważność. Inną motywacją, poza statystyczną ważnością, dlaczego interesujący jest ten współczynnik, jest fakt, że poszukiwane reguły powinny spełniać pewne wymaganie co do wysokości wartości *sup* z powodów biznesowych. Jeśli poziom pokrycia nie jest wystarczająco wysoki, to oznacza to, iż reguła nie jest warta brania pod uwagę, bądź jest ona mniej preferowana (może być rozpatrzona później) [AIS93].

Literatura

[Agr94] Rakesh Agrawal. Quest: a project on database mining. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1994.

[AIS93] Rakesh Agrawal, Tomasz Imielinski, Arun Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, strony 207–216, 1993.

[EN05] Ramez Elmasri, Shamkand B. NAvathe. *Wprowadzenie do systemów baz danych.* Addison-Wesley, Reading, MA, USA, 2005.

[FPSSU96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

Instytut Informatyki, Wydział Informatyki i Zarządzania
Politechnika Poznańska

Skład przy użyciu systemu LaTeX.

BibTeX:

```
@mastersthesis{ key,
    author = "Tomasz Kujawa",
    title = "{Równoległe odkrywanie reguł asocjacyjnyh zaimplementowane na procesory graficzne}",
    school = "Poznan University of Technology",
    address = "Pozna{\'n}, Poland",
    year = "2011",
}
```