POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

KOMPUTEROWE WSPOMAGANIE ZARZĄDZANIA I PROCESÓW DECYZYJNYCH W GOSPODARCE

pod redakcją:
Jana Studzińskiego
Ludosława Drelichowskiego
Olgierda Hryniewicza

KOMPUTEROWE WSPOMAGANIE ZARZĄDZANIA I PROCESÓW DECYZYJNYCH W GOSPODARCE

Polska Akademia Nauk • Instytut Badań Systemowych

Seria: BADANIA SYSTEMOWE
tom 31

Redaktor naukowy:

Prof. dr hab. Jakub Gutenbaum

Warszawa 2002

KOMPUTEROWE WSPOMAGANIE ZARZĄDZANIA I PROCESÓW DECYZYJNYCH W GOSPODARCE

pod redakcją
Jana Studzińskiego, Ludosława Drelichowskiego
i Olgierda Hryniewicza

Książka zawiera wybór artykułów poświęconych omówieniu aktualnego stanu badań w kraju w zakresie rozwoju i zastosowań technologii, modeli i systemów informatycznych w gospodarce narodowej.

Recenzenci artykułów:

Prof. dr hab. inż. Olgierd Hryniewicz
Prof. dr hab. inż. Janusz Kacprzyk
Dr inż. Lech Kruś
Dr inż. Edward Michalewski
Prof. dr hab. inż. Andrzej Straszak
Dr inż. Jan Studzinski
Dr inż. Sławomir Zadrożny

Wydawca: **Instytut Badań Systemowych PAN**
ul. Newelska 6 01-447 Warszawa

Redakcja: Dział Informacji Naukowej i Wydawnictw IBS PAN
tel. 837-68-22
Barbara Kotuszewska

Druk: Zakład Poligraficzny Urzędu Statystycznego w Bydgoszczy
Nakład 200 egz. ark. wyd. 23,5 ark. druk. 20,0

ISBN 83-85847-73-1
ISSN 0208-8028

Rozdział 3

Metody i algorytmy obliczeniowe w systemach wspomagania decyzji

WYSZUKIWANIE PODOBNYCH SZEREGÓW CZASOWYCH METODĄ OPISU SYMBOLICZNEGO

Krzysztof Kania
Akademia Ekonomiczna w Katowicach
<kkania@ae.katowice.pl>

Finding similarity of time-series is an important issue in temporal data mining. It may be useful in many areas including market and economy analysis, customer behavior, prediction of natural disasters, diagnosis and monitoring of processes etc. Finding objects with similar behavior can be also used for explanations in expert systems. Works focus on defining similarity measures for time-series and sequences and building fast and efficient algorithms for analyzing time-series and sequences. This article presents a method for finding similar time-series based on symbolic description.

Keywords: time-series similarity, data mining, symbolic methods.

1. Tryby analizy danych

Praca modułów analizy danych temporalnych może odbywać się w dwóch trybach: off-line oraz on-line (zob. np. Cooley 1999). Wybór trybu zależy od celu badania. W trybie off-line odkrywanie i aplikacja wiedzy jest zwykle procesem długotrwałym lecz dokładnym i kompletnym. Ograniczenia czasowe w takim przypadku są bardzo słabe. Odkryta wiedza ma często charakter stosunkowo trwały i może mieć wpływ na wszystkie elementy funkcjonowania organizacji. W tym trybie udział wiedzy wstępnej może być ograniczony do wskazania celu drążenia oraz zasad, według których drążenie ma się odbywać. Na tym etapie realizowane są takie zadania jak: ekstrakcja reguł, grupowanie, wyszukiwanie częstych sekwencji itd. W tym trybie pracy najszersze zastosowanie mają procedury oparte o różne warianty algorytmu Apriori zaproponowanego po raz pierwszy w (Agraval 1994).

Tryb on-line przewidywany jest do szybkiej reakcji na zaistniałą sytuację. W takich przypadkach konieczne jest przyjęcie dużego udziału wiedzy wstępnej (przede wszystkim tej uzyskanej w procesie drążenia danych w trybie off-line oraz wiedzy fundamentalnej) oraz zastosowanie szybkich, chociaż przybliżonych algorytmów. Ten tryb pracy systemu wykorzystywany jest przede wszystkim do analizy i klasyfikowania kolejnych pojawiających się przypadków. Na podstawie wyników przeprowadzonej analizy realizowane są działania dostosowujące bieżącą

pracę systemu do pojawiających się sytuacji poprzez wspieraną wiedzą reakcję na odbierane sygnały. Ze względu na potrzebę szybkiej analizy bardzo dużych ilości danych badacze podejmują wysiłki zmierzające do zoptymalizowania algorytmów analizy SzC. Badania koncentrują się na dwóch zagadnieniach:

- przekształceniu wartości SzC zmierzającym do przyspieszenia obliczeń (transformacja i segmentacja) oraz budowy odpowiedniego indeksu opisującego SzC, który powinien zapewniać (Faloutsos 1994): szybsze obliczenia niż przeszukiwanie sekwencyjne, działanie w warunkach ograniczonych zasobów obliczeniowych, odpowiedzi na zapytania o szeregi różnej długości, możliwość usuwania i dodawania danych bez konieczności przebudowy indeksu, poprawność tzn. nie wykluczać najlepszych dopasowań.
- dokonaniu opisu numerycznego lub symbolicznego umożliwiającego selekcję kandydatów poprzez poszukiwanie kandydatów do badania lub eliminację kandydatów nie rokujących nadziei na podobieństwo.

Szereg czasowy można interpretować jako punkt w wielowymiarowej przestrzeni. Ilość wymiarów odpowiada długości SzC. Problem znalezienia najlepiej dopasowanych SzC można zatem sprowadzić do problemu najbliżej leżącego sąsiada w wielowymiarowej przestrzeni. Znane i dostępne są bardzo wydajne algorytmy indeksowania na podstawie odległości pomiędzy punktami. Ponieważ jednak sprawnie funkcjonują one tylko w ograniczonej liczbie wymiarów pojawiły się próby takiej transformacji SzC, która zachowując odległość pomiędzy szeregami, sprowadziłaby je do jak najmniejszej liczby wymiarów. Jedną z najbardziej udanych prób podniesienia wydajności indeksowania opartą na tych założeniach jest przekształcenie euklidesowej odległości między SzC w Dyskretnej Transformacji Fouriera na współczynniki częstotliwości i użycie kilku pierwszych z nich do zbudowania indeksu (Agraval 1993). Również na zasadzie redukcji ilości wymiarów opiera się optymalizacja, którą zaproponowano w (Keogh 2000). Skonstruowano ją na założeniu, że można dokonać przybliżonego opisu sekwencji danych poprzez segmentację sekwencji na określoną ilość części, obliczenie średniej arytmetycznej dla każdego segmentu i przyjęcie zapisanych średnich za podstawę budowy indeksu. W tym wypadku ilość wymiarów zostaje zredukowana do ilości równej ilości segmentów.

Proponowaną w niniejszym artykule metodę można zakwalifikować do grupy metod, które umożliwiają przyspieszenie badania numerycznych SzC poprzez selekcję kandydatów. Ze względu na fakt, że symboliczny opis SzC jest zawsze opisem przybliżonym proponowana metoda jest przewidywana do pracy w trybie on-line. Metoda ta:

- jest niezależna od stosowanej miary podobieństwa ponieważ opiera się na zasadzie badania kandydatów i odrzucaniu kandydatów nie rokujących nadziei na podobieństwo,

- jest oparta o indeks symboliczny, który można przechowywać łatwo w bazach danych,
- nie wyklucza połączenia z innymi metodami,
- umożliwia użycie języka opisu wzorców i formacji definiowanych przez użytkownika i włączenia go do algorytmu wyszukiwania.

2. Symboliczny opis szeregu czasowego - założenia metody

Ogólnie algorytm poszukiwania podobnych SzC oparty o przekształcenie symboliczne można opisać następująco:

1. zbuduj krótki opis badanych SzC,
2. porównaj zbudowane charakterystyki,
3. jeżeli są one podobne
   zakwalifikuj SzC do dalszych obliczeń,
   w przeciwnym wypadku
   odrzuć.

Na pierwszy rzut oka może wydać się nieracjonalne zastępowanie jednej miary podobieństwa inną. Cały sens takiej operacji polega na tym, że porównanie krótkich opisów jest szybsze niż wykonywanie obliczeń na całych SzC. Proponowana metoda opiera się na oczywistej obserwacji, że jest mało prawdopodobne, aby dwa SzC były do siebie podobne, jeżeli przebiegi poszczególnych jego części są od siebie różne. Z drugiej strony można powiedzieć, że jeżeli przebiegi poszczególnych części szeregu są do siebie podobne to wskazuje to, że całe szeregi mogą być do siebie podobne w stopniu interesującym użytkownika. Możliwe jest zatem przerwanie procesu obliczeniowego, jeżeli tylko okaże się, że odpowiadające sobie fragmenty szeregów (badanego i wzorca) są do siebie niepodobne. Proponowana metoda adaptuje również idee symbolicznego opisu SzC zaprezentowane w (Agraval 1995) oraz (Hebrail 2000).

Do realizacji metody potrzebne są:

- funkcja budowania symbolicznych opisów SzC oraz
- metoda badania podobieństwa tych opisów.

Do budowania opisu SzC przyjmujemy funkcję *f*, która jest oparta o badanie relatywnych zmian wartości w szeregu. Przyjęcie takiej postaci funkcji odpowiada definicji, według której przebieg dwóch zjawisk jest tym bardziej podobny im różnica między przyrostami względnymi jest mniejsza a kierunki zmian zgodniejsze. Możliwość zaadoptowania innych definicji podobieństwa zależy tylko od możliwości zdefiniowania na jej podstawie odpowiedniej funkcji. W zależności od potrzeb możliwe jest zastosowanie opisu składającego się ze zmiennej ilości znaków alfabetu.

Przykładowo:

$$f_3(\Delta x)=\begin{cases}'+' \Leftrightarrow \Delta x>\varepsilon\\ '0' \Leftrightarrow |\Delta x|\le\varepsilon\\ '-' \Leftrightarrow \Delta x<-\varepsilon\end{cases} \qquad f_5(\Delta x)=\begin{cases}'/' \Leftrightarrow \Delta x\ge 2\varepsilon\\ '+' \Leftrightarrow 2\varepsilon>\Delta x>\varepsilon\\ '0' \Leftrightarrow |\Delta x|\le\varepsilon\\ '-' \Leftrightarrow -2\varepsilon<\Delta x<-\varepsilon\\ '\backslash' \Leftrightarrow \Delta x\le -2\varepsilon\end{cases}$$

gdzie: ε jest wartością arbitralnie określoną przez użytkownika oraz,

$$\Delta x=\frac{x_i}{x_j}-1 \quad i,j\in N,\ j>i.$$

Przyjęte wartości ε oraz wielkości przedziałów odpowiadające poszczególnym znakom ściśle zależą od badanej dziedziny. Od przyjętych przedziałów zależy też interpretacja znaków np. przedział $[-\varepsilon,\varepsilon]$ wyznacza obszar, w którym zmiany wartości nie są istotne dla użytkownika, '+'-wzrost, '/'- duży wzrost itd. Możliwe jest także uzależnienie wartości ε od wartości wyrażenia $k=j-i$. Podaje się wtedy ε_p (początkowe) oraz ε_k (końcowe) oraz sposób obliczenia ε dla konkretnego przypadku. Przykładowo, gdy chcemy, aby dłuższy czas pomiędzy elementami szeregu odpowiadał większemu ε można przyjąć:

$$\varepsilon=\varepsilon_p+\frac{\varepsilon_k-\varepsilon_p}{n}(j-1), \qquad \text{gdzie } n \text{ – długość badanego szeregu.}$$

Gdy na podstawie funkcji f wygenerowane zostaną już opisy symboliczne SzC należy sprawdzić, czy są one na tyle podobne, że warto zakwalifikować same szeregi do dalszego szczegółowego badania czy nie. Konieczne jest w tym celu określenie jakiejś metody porównywania opisów. Do jej skonstruowania przyjęto dwa założenia:

- jest określona różnica pomiędzy znakami alfabetu,
- możliwe jest nałożenie wag na poszczególne znaki opisu symbolicznego.

W zależności od preferencji można określić inne różnice między znakami. Dla alfabetu 3 i 5 znakowego mogą one być następujące:

3	-	0	+
-	0	½	1
0		0	½
+			0

5	\	-	0	+	/
\	0	¼	½	¾	1
-		0	¼	½	¾
0			0	¼	½
+				0	¼
/					0

Z kolei nałożenie wag na poszczególne części opisu odzwierciedla fakt, że dla użytkownika różne części opisu mają różną wartość, co równocześnie daje szersze możliwości modelowania preferencji użytkownika. Na przykład waga może wzrastać wraz z długością badanego fragmentu, co jest odbiciem sytuacji, gdy zmiany w dłuższych okresach czasu są bardziej znaczące dla użytkownika niż zmiany w okresach krótszych, w podobny sposób można wymodelować na przykład to, że zmiany nowsze są istotniejsze niż starsze. Na tej podstawie można już zdefiniować miarę podobieństwa opisów:

$$\frac{\sum(\text{różnica pomiędzy znakami}\times\text{waga znaku})}{\sum(\text{różnica maksymalna między znakami}\times\text{waga znaku})}*100\%$$

Miara przyjmuje wartości z przedziału [0;1] i wynosi 0 dla szeregów, o identycznych opisach, 1 dla szeregów o opisach całkowicie różnych i jest w dalszej części pracy wyrażana w procentach. Do dalszych badań kwalifikowane są te SzC, dla których miara jest niższa niż założona wartość.

3. Przykład wykorzystania metody

Samo zbudowanie indeksu jest zadaniem, na które nie nakłada się ostrych ograniczeń czasowych, natomiast jego budowa musi umożliwiać szybkie działanie w przyszłości i łatwą aktualizację. W proponowanym rozwiązaniu z każdym elementem szeregu jest skojarzony łańcuch opisujący zmianę jego wartości w stosunku do *n* kolejnych elementów SzC. *J*-ty znak opisu dla *i*-tego elementu SzC jest obliczany zgodnie z zadaną funkcją opisu dla *j=1,2,...,n*. Wartość *n* określa maksymalną długość badanych szeregów, która musi być określona z góry. Opis konkretnego fragmentu SzC jest tworzony z odpowiednich znaków przechowywanego opisu. Przyspieszenie polega na zastąpieniu tworzenia opisu poprzez obliczanie przez tworzenie poprzez wybieranie fragmentów opisów przechowywanych w indeksie.

Tablica 1: Przykład indeksu symbolicznego

Kolejne wartości	Opis symboliczny n=16	wartości c.d.	opis symboliczny n=16
4989,95	00+++++0++++++++	5156,86	00+00-0-0000
4983,09	++++++++++/+++++	5184,32	0000-0-0000
5023,55	0000+00+++++++++	5174,92	+00-0-0000
5041,61	000+00+++++++++0	5216,47	00\-\----
5048,84	00+00+++++++++00	5182,15	0\-\--00
5070,88	0000++++0+0+0000	5176,73	\-\--00
5078,10	000++++0+0+00000	5075,21	000000
5105,56	000++0000+000000	5109,89	-0000
5074,49	0++/++++++000000	5059,32	0000
5087,13	++/+++++00000000	5096,53	000
5139,52	0+00000000000000	5097,97	00
5177,45	00000000-0-0000	5110,26	0
5199,13	-000000-0-0000	5105,92	
5159,39	000000-0-0000-		

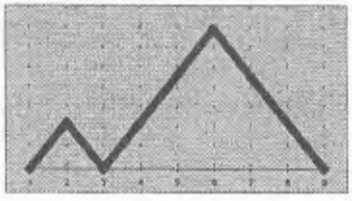
Znaczenie symboli zacienionych objaśniono w tekście.

Budowa indeksu jest oparta o wybraną funkcję opisu symbolicznego dla zadanej maksymalnej długości opisu. W tablicy 1 przedstawiono przykład indeksu zbudowanego dla funkcji f_5 z parametrami ε_p=0,75% ε_k=3% i n=16.

Modyfikacja indeksu jest prosta - dodanie kolejnego elementu do SzC powoduje konieczność uzupełnienia *n* końcowych opisów przez obliczenie stosownych przyrostów i dopisanie do nich kolejnego znaku. Indeks daje się łatwo przechowywać w bazie danych – zajmuje jedno pole znakowe o długości takiej jak przyjęta długość opisu. Indeks wykorzystujemy do zbudowania symbolicznego opisu rzeczywistego lub zbudowanego przez użytkownika wzorca. Ponadto użytkownik definiuje odpowiednie wagi. Zatem symboliczny opis ciągu to zbiór trójek:

<miejsce początkowe M_p, miejsce końcowe M_k, waga>.

Przykładowo poszukujemy w szeregu formacji z rysunku 1.



Rys. 1. Przebieg przykładowego wzorca

Wskazując na punkty charakterystyczne formacji można zdefiniować opis szeregu i przyłożyć do znaków opisu odpowiednie wagi. W tabeli 2 zawarto opis wzorca z rysunku 1, wagi wprost proporcjonalne do długości fragmentu szeregu.

Tabela 2: Opis przykładowego wzorca

M_p	M_k	waga	znak	Interpretacja
1	2	1	+	między 1 i 2 wartością nastąpił wzrost
1	3	2	0	nie występuje istotna różnica między wartością 1 i 3
1	9	9	0	nie występuje istotna różnica między wartością 1 i 9
2	3	1	-	między 1 i 2 wartością nastąpił spadek
3	6	3	/	między 3 i 6 wartością nastąpił znaczny wzrost
6	9	3	\	między 6 i 9 wartością nastąpił znaczny spadek

Ostatecznie otrzymujemy opis: `+00-/\`. Kolejność znaków w opisie jest istotna tylko dla efektywności odczytu z indeksu – w celu przyspieszenia obliczeń należy rozpocząć od znalezienia znaków o najwyższej wadze – wtedy w przypadku wystąpienia istotnej różnicy możliwe jest natychmiastowe przerwanie obliczeń.

Do zbudowania opisu dla kolejnych fragmentów przeszukiwanego szeregu należy pobierać odpowiednie znaki z indeksu. Przykładowo dla podciągu rozpoczynającego się od trzeciej wartości szeregu z tabeli 1 zostanie wygenerowany opis (zob. wartości zacieniowane): `00+0+0`. Uzyskane opisy należy porównać. Potrzebne dane zestawiono w tablicy 3, przyjmując wartości z przykładów.

Tablica 3: Przykładowe obliczenia do porównania opisów

Znaki opisu szeregu 1	+	0	0	-	/	\
Znaki opisu szeregu 2	0	0	+	0	+	0
Różnica między znakami	0,25	0,00	0,25	0,25	0,25	0,5
Waga	1	2	9	1	3	3
Iloczyn wagi i różnicy	0,25	0,00	2,25	0,25	0,75	0,75

Suma wartości ostatniego wiersza wynosi 4,25, suma iloczynów wag i największej różnicy między znakami (jedynka z tabeli różnic znaków) wynosi 19. Zatem wartość miary porównywania opisów wynosi 4,25/19=22,37%. Jeśli jest to

mniej niż przyjęta wartość progowa uznajemy, że opisy są na tyle podobne, że warto poddać je dokładnemu badaniu. W przeciwnym wypadku odrzucamy.

Samo użycie wartości względnych do budowy opisu rozwiązuje problem analizy szeregu w różnej skali pionowej. W przypadku, gdy użytkownik jest zainteresowany znalezieniem wzorca w różnej skali poziomej należy rozciągnąć lub ścieśnić wzorzec proporcjonalnie zwiększając lub zmniejszając odpowiednie wartości M_p i M_k.

4. Podsumowanie

Zaprezentowana metoda umożliwia elastyczne badanie SzC przez użytkowników oraz przyspieszenie wykonywanych działań. Użycie indeksu wprawdzie znacznie zwiększa ilość miejsca potrzebnego do przechowania szeregu jednak sposób budowy indeksu umożliwia łatwe przechowanie go w relacyjnych bazach oraz łatwe przeszukiwanie. Do ograniczeń podejścia należy zaliczyć konieczność zbudowania symbolicznego opisu wzorca, co z jednej strony może być trudne dla użytkowników bez przeszkolenia, a z drugiej uniemożliwiać automatyzację badania pojawiających się przypadków. Rozwiązaniem tego problemu może być zbudowanie modułu automatycznej budowy opisów na podstawie punktów charakterystycznych szeregu oraz opracowanie odpowiedniego interfejsu graficznego. Ponadto ze względu na sam charakter opisu symbolicznego znalezione rozwiązanie należy traktować jako przybliżone.

Literatura

Agraval R., Faloutsos Ch., Swami A. (1993): Efficient Similarity Search in Sequence Databases, Proc of the 4th Intl. Conf. On Foundation of Data Organization and Algorithms, Chicago.

Agraval R., Psaila G., Wimmers E., Zait M. (1995): Querying Shapes of Histories, Proc. Int. Conference Very Large Data Bases, Zurich, Switzerland, 502-514.

Agraval R., Srikant R. (1994): Fast algorithms for mining association rules, Proc. Int. Conference Very Large Data Bases, Santiago, Chile, 487-499.

Cooley R., Tan P., Srivastava J. (1999): Discovery of Interesting Usage Patterns from Web Data, Technical Report TR99-022, University of Minnesota.

Faloutsos C., Ranganathan M., Manolopoulos Y. (1994): Fast subsequence matching in time-series databases, Proc. ACM SIGMOD Conf. Minneapolis, 419-429.

Hebrail G., Huguney B., 2000: Symbolic Representation of Long Time-Series, Proc. of Workshop Symbolic Data Analysis PKDD.

Keogh E., Pazzani M. (2000): A simple dimensionality reduction technique for fast similarity search in large time series databases, 4th Pacific-Asia Conf. On Knowledge Discovery and Data Mining, Kyoto.

ISSN 0208-8028
ISBN 83-85847-73-1

W celu uzyskania bliższych informacji i zakupu dodatkowych egzemplarzy
prosimy o kontakt z Instytutem Badań Systemowych PAN
ul. Newelska 6, 01-447 Warszawa
tel. 837-35-78 w. 241 e-mail: bibliote@ibspan.waw.pl